BRIDGEHEALTH
BRidging Information and Data Generation for Evidence-based Health Policy and Research

WP8

Platform for population-based registries

Task 2 – Deliverable 8.2/M35

Blueprint of open source software platform for population-based chronic disease registries

Final Report
October 2017

UNIVERSITA' DEGLI STUDI DI ROMA TorVergata
UNIVERSITY OF SURREY
HRS
EuBIROD NETWORK

This report is Deliverable D8.2 of “WP8. Platform for population-based registries (BRIDGE-HEALTH Project”, co-funded by DG-SANCO, European Commission, 2015 (G.A. 664691).

This report is composed of 50 pages.

Scientific Coordinator: Prof.Herman Van Oyen (Belgian Institute of Public Health)
WP Leader: Simona Giampaoli (National Institute of Health, Italy).
Task Leaders: Prof.Massimo Massi Benedetti (HIRS, Italy), Prof. Luigi Uccioli (UNITORV, Italy), Prof. Fabrizio Carinci (University of Surrey, UK).

The Bridge Health Task 8.2 Study Group:

Doggen K, Inst. Scient. Santé Pub. WIV, Brussels, Belgium
Poljicanin T, Metelko Z, Pristas I, National Institute of Public Health, Croatia
Traynor V, Olympios G, Ministry of Health, Larnaca, Cyprus
Jecht M, Havelhoe Hospital - Gabris L, Scheidt-Nave C, Robert Koch Institute, Germany
Janos S, Nagy A, Adany R, University of Debrecen, Hungary
Zucker I, Israel Center for Disease Control, Israel
Di Iorio CT, Serectrix snc; Gualdi S, Uccioli L, University Tor Vergata; M.Monesi, National Association of Diabetologists; Massi Benedetti M, HIRS, Italy
Ramuse E, Golosujeva L, Lepiksone J, Centre for Disease Prevention and Control, Latvia
Azzopardi J, Fava S, University of Malta, Malta
Polanska J, Jarosz-Chobot P, Silesian University of Technology, Katowice, Poland
Pruna S, Telemedica Consulting, Romania
Stotl I, University Children’s Hospital, Ljubljana, Slovenia
S.Gudbjornsdottir, National Diabetes Register, Sweden
Cunningham S, Brillante M, University of Dundee, Scotland, United Kingdom
Carinci F, De Lusignan S, University of Surrey, England, United Kingdom

WP Leader: National Institute of Health, Italy
Task Leader: University Tor Vergata, Italy

Compiled for the Study Group by:
Fabrizio Carinci, Visiting Professor, University of Surrey, UK

Citation
Carinci F, Gualdi S, Di Iorio CT and Massi Benedetti M on behalf of Bridge Health Task 8.2 Study Group, Blueprint of open source software platform for population-based chronic disease registries, Bridge Health Consortium, October 2017.

CHAPTERS

CHAPTER 1

Population-based registries for chronic diseases

1.1. Definitions

According to the guidelines of the Joint Action PARENT: a patient registry is „... an organized system that collects, analyses, and disseminates the data and information on a group of people defined by a particular disease, condition, exposure, or health-related service, and that serves a predetermined scientific, clinical or/and public health (policy) purposes“ (PARENT Consortium 2014).

Disease or condition registries „..are defined by patients having the same diagnosis, such as cystic fibrosis or heart failure, or the same group of conditions such as disability.“. Here EUBIROD is explicitly mentioned as: *„..as an example of an EU project/initiative concerning improving disease registries in terms of defining purposes, legal context, semantic and technical aspects..Overall, EUBIROD can serve as a good example and model to be re-used for other chronic diseases as well*“.

A Population Registry *„... is a registry that intends to cover all residents in a given geographic area within a given time period. The coverage of the specific registry may, however, be incomplete, but it is nevertheless a population registry if the aim is to include all the individuals in the target population. A population is defined by geographical boundaries, but usually only residents (or citizens)within a given time period are included in the definition.*“.

A Population-based registry should be used *”...when all persons with a given trait, exposure or event, are intended to be included in the registry. If the registry includes everyone in the population (even the oldest), it becomes a population registry. Intention rather than performance defines the terms. A population-based disease registry aims at including everyone with the disease in the population, be it self-reported, clinically diagnosed or detected at screening. Population and population-based registries may be further classified as of good or bad quality depending on coverage or other characteristics*“.

1.2. The BIRO experience

The conceptual framework of the BIRO system (2005-2008, www.biro-project.eu) endorsed by the EUBIROD network (2008-2012, www.eubirod.eu) emerged as a common infrastructure modelled on top of existing implementations in Europe (BIRO Consortium 2009).

According to this approach (BIRO Consortium 2009, Di Iorio et al. 2009, Carinci et al. 2010a, Di Iorio et al. 2013b), the statistical analysis of individual data should only takes place at the local level, while in all other instances information is shared and processed only in the form of anonymous aggregate data.

This infrastructure allows participating centres to generate a report for internal use ("local"), while creating a stream of finely tuned micro-aggregates, which can be transmitted

to a higher level in a hierarchy of servers (regional, national, European). At each of these levels, data are stored and analysed only in aggregate format, thus with limited burden and privacy risk for the data custodians. At each node of the server pipeline, users can produce cumulative reports for all the lower levels in the hierarchy, using the same software ("central") and the same procedure adopted at the local level. Moreover, if a unique centralised database is formed including all the micro-aggregate data with proper and accurate tagging of origin and time reference, then all combinations of comparisons at all levels become possible.

Such information infrastructure does not only allow maximum safety in terms of privacy protection, but favours resource optimization at all levels for two fundamental reasons: a) the overload for data linkage and processing is distributed across sites, without the need to create a massive cumulative database, e.g. a European register. b) calculation of indicators is simplified, as they can create out of partial cumulative tables includinf case counts for well-defined categories (eg. number of males older than 65 years with over 50% examinations performed out of those expected for the specific individual conditions).

Obviously, there are drawbacks as not all the analyses are possible without using individual data. However, for targeted studies on longitudinal cohorts it is always possible to pursue ethical approval, to extract and exchange personal records. With a system e.g. BIRO in place, even these tasks, given the level of standardisation imposed to the network, would become easier to carry out.

A system e.g. BIRO makes possible to create national reports using the most updated data, with a total processing time that is not much longer than what is necessary for the largest local health authority or region.

A regional population-based disease register may represent the most effective building block to operate a system e.g. BIRO under the best conditions, as it allows has computing an accurate numerator (eg the absolute number of major amputations in people with diabetes) on top of a well defined, area-based denominator (eg the total number of people with diabetes in a specific territorial area eg catchment area, province or region).

In this way, collaborative networks of population-based registries can transform local data sources into powerful engines of actionable information for policy makers, health professionals and citizens at all levels.

1.3. Task 8.2 of Bridge Health

The main aim of the task is '*to maintain and strengthen the implementation of population based registries for chronic diseases through the standardization of methodologies for producing standardized EU-wide indicators, taking selected clinical conditions as test cases for a new 'platform for population based registries*'.

A specific objective is the provision of privacy-enhanced software for statistical analysis, data exchange, and automated calculation of indicators, locally and at EU level.

This task will take advantage of the continuing EUBIROD network of registers, coordinated by HIRS, to make further progress and deliver results across different diseases. The further

development of open source software for data management, statistical analysis and automated delivery of indicators is planned to facilitate operations through a user friendly interface that will further enable data custodians to produce local reports and transmit data towards a central location, for the routine production of EU indicators (e.g. ECHI shortlist).

The level of compliance of the whole process with privacy and data protection rules will be assessed by a targeted evaluation method, shared with other relevant workpackages. The development of technical manuals will allow strengthening the use of software by personnel involved in data processing of population-based registers.

The deliverables of Task 8.2 include:

- Month 18:
  - D8.2 Blueprint of open source platform for population-based chronic disease registers (draft)
  - D8.3 Manual of technical specifications for users and programmers (draft)
- Month 30:
  - D8.5 Blueprint of open source software platform for population-based chronic disease registers (final)
  - D8.6 Manual of technical specifications for users and programmers (final)

The work will be conducted by personnel at the University of Tor Vergata and the University of Surrey, supported by experts of the EUBIROD network gathering for two annual meetings during the 30 months of duration of the project.

CHAPTER 2

Why open source software for disease registries

2.1. The cathedral and the bazaar

The "Cathedral and the Bazaar" is a book published by Eric Raymond in the late 90s (Raymond, ES 1999). It includes an early presentation of open source software whose principles are still valid, particularly in the health sector.

The starting point of this book was the concept of Cathedral represented in a book published in 1975 by Fred Brooks, "*The Mythical Man-Month*", which states that "*conceptual integrity is the most important consideration in system design*". In the software industry, this translates into the paradigm where the stronger products emerge from solid design produced by one or more talented engineers, who would coordinate its implementation in a top-down approach.

In opposing the metaphor of a "bazaar", Raymond presented open source as an alternative approach, where eventually such a design does not it even exist, at least in its early inception. In his words "*cooperative software development effectively overturns Brooks' Law leading to unprecedented levels of reliability and quality on individual projects*".

Raymond's review of the open source phenomenon was highly effective and influential. It opened the way to the release of fundamental source code for the development of internet services e.g. the early Netscape Communicator. It resulted into a boost for the early work of Richard Stallman on Open software and Linus Torvalds on Free Unix Systems (Linux).

However, more than just being prophetic, these words described an underlying revolution that changed how communities of professionals interconnect and grow projects that use software *as a means to an end,* rather than as a business objective.

The health sector benefited immensely from these developments.

Today, open source software has become mainstream, so that the many alternatives available are not only exceptionally performing - rivalling commercial proprietary alternatives – but they can be effectively linked together to create comprehensive applications.

Users can access Java programming languages, together with database management systems e.g. PostgreSQL/MySQL and powerful statistical languages e.g. R, with well over 10,000 packages available on a shared repository.

Proprietary software had to revise own business model, to concentrate on the ease of use and more proactive continued assistance, which obviously is not possible to provide at the individual level to a specific client.

On the other hand, the same models are also applicable to open source software, as the tools can be downloaded and installed for free, but if they have to be customised, then would require specific contracts for the provision of services.

Therefore, we could safely state that open source and proprietary software today share an

equal ground, very differently from the time of publication of Raymond's book.

Today, there is a plethora of open source licenses, and having the source code open, does not mean that the software cannot be inherently “commercial”. However, we would use this concept with clear reference to the original idea of “community effort”, which can be translated in the scope producing the blueprint for population-based registries.

How would the above development translate into a solid application of the kind we are proposing here?

There are two different areas in which the choice of software is critical:

1. at the point of data collection
2. data analysis, reporting and exchange.

As far as data entry and storage is concerned, this is very likely an area where commercial software still has an edge over open source. That is because the needs may be easier to standardize and the commercial companies may distribute the cost of developing highly customisable applications over a multitude of users. However, this can also depend from the local environment: where projects emerge with a bottom-up approach, it might be more likely that open source is used to create own database platforms.

On the other hand, the analytical needs may be much more difficult to standardise towards common models, and even the best industrial software (e.g. SAS) may require a significant effort to be programmed and adapted. Moreover, information transmission and security protocols have been very strongly supported by the open source community. A great advantage is also offered by having the possibility to access all updates and new methods that are immediately implemented as open source by the academic sector. Last, but not least (see next section) the European Commission actively contributes to open source development, which translates into increased adoption in relevant EU projects.

As a result, for the second order of needs, it is easier to configure an open source product, taking advantage from the continuously evolving existing libraries.

So, what should be the ideal approach for the construction of a platform in Task 8.2 of Bridge Health? As an early supporter of open source software, the EUBIROD network reinforces this vision as the best approach possible for the development of a comprehensive system. That does not preclude collaboration with proprietary software that may be interconnected to the platform in various ways.

Obviously, there are strengths and weaknesses, risks and opportunities of this approach, which are shared with open source in general. For a more detailed discussion, the interested user may refer to the extensive literature available on this topic.

Above all, we believe that the principle of the *bazaar* has practical conveniences over the *cathedral* in the design of a cohesive platform (**Figure 2.1**) for the following reasons:

- avoids relying upon a single product, which may eventually fail without a clear plan for succession. This has been quite frequent in the health sector, particularly on international cooperation.
- most appropriate for global development (easier access for low and middle income countries)

- fosters the development of a growing community
- allows making the model sustainable and open to future developments
- impedes incorporation into commercial entities or proprietary software
- encourages new investments from international organizations

The issue of licensing is extremely relevant to realise the above advantages.

Figure 2.1. The “cathedral and the bazaar”

2.2. Open source licenses

A plethora of licenses are available today for open source software. Although the number of approaches may be classified in various ways, the most important feature is whether the derived source can be incorporated into proprietary software and become by all means open source code.

This is not the case with the original definition endorsed by the General Public License (GPL), which is in fact frequently described as a “viral” approach. In practical terms, under the GPL 2.0, all derivative works of the software and subsequent versions down the chain must be licensed and distributed on the same terms as the original software.

In this way, if a developer wishes to create a new program, if he/she wants it to be of the greatest possible use to the public, the best way to achieve this is to make it free software which everyone can redistribute and change under these terms.

Source code subject to the GPL permanently remains subject to GPL. This permanent nature of the GPL, as intended by the authors of the GPL, constrains the options available to developers building on GPL software in creating, distributing or commercializing products using existing GPL source code. There are also other potential challenges faced by developers, for instance in determining when software developed for a GPL software platform is considered a derivative work that is subject to the GPL.

However, charging fees for system setup, system management, support, maintenance and other related services is permitted under the GPL.

The license states that *“This General Public License does not permit incorporating your program into proprietary programs. If your program is a subroutine library, you may consider it*

more useful to permit linking proprietary applications with the library. If this is what you want to do, use the GNU Lesser General Public License instead of this License".

The "LGPL 2.0/3.0 License" is intended to permit developers of non-free programs to use free libraries, while preserving their freedom as users of such programs to change the free libraries that are incorporated in them. In this way, if a programmer develops a new library, and he/she wants it to be of the greatest possible use to the public, it can be licensed under the LGPL to redistribute and change.

Other licenses, e.g. Mozilla, Apache, Creative Commons etc are less restrictive in terms of the modifications and redistribution of programs under different terms.

2.3. Strategy of the European Commission and the EUPL

The European Commission has released a new strategy 2014-2017 for open source software.

The strategy has been specifically released to ensure that:

- open source and proprietary software are assessed on an equal basis, being both evaluated on the basis of total cost of ownership, including exit costs.
- products support recognised, well-documented and preferably open technical specifications that can be freely adopted, implemented and extended with a view towards interoperability and use of well-established standards
- participation of open source communities to EC software is increased (particularly through DIGIT)
- legal advise will be provided to deal with intellectual property issues
- open source software is the preferred choice for new information systems supported by the Commission
- state of the art open source software is ensured particularly in the areas of security and e-Governance
- software produced by the Commission will be open sourced and published on the Joinup platform with the European Union Public License (EUPL)

As shown by this last point the EUPL, being the first open source licence ever released by an international governing body, is an important building block in this strategy. Its first version was approved on 9/1/2007. Its aim is to clarify issues around the governance of other licences, above all the GPL, which refers to the US legislation, while the new one takes due account of European Union Law.

The EUPL is also available into 22 official languages of the European Union, so it assures to conform to the existing copyright laws in each Member State. It also authorizes to be re-released under different licences, particularly the GPL 2.0.

With these developments in mind, it is straightforward to support the EUPL as the license of choice for the development of an open source platform for population-based registries.

CHAPTER 3

The problem of bias in disease registers

3.1. Issues in the design of disease registers

Disease registers are frequently built as a cohort study: at a certain date an event, e.g. a visit, examination or even birth for certain diseases, a diagnosis triggers the registration into a cohort. The subject enters a database in which all personal characteristics (e.g. age, sex), clinical measurements (e.g. blood pressure), processes (e.g. visits, prescriptions, etc), intermediate (e.g. HbA1c) and terminal outcomes (e.g. renal failure, amputations, death) are regularly updated.

The extent of the coverage of the register from the initial recording to exit from the database (for migration or death) determines the completeness of information available for routine care or research, as well as the capacity of the system to produce unbiased indicators for public health monitoring.

Population-based registers take carefully into account the relation between individual data and the total cohort registered in the database, estimating results that can be evaluated on a person basis, rather than being provider-oriented, for a specified catchment area.

As systems evolved, it has become increasingly possible to integrate information from different sources through unique IDs shared across a network of centres, e.g. hospitals, outpatient/specialist clinics, primary care and the government. In this way, the performance of health care for people with a chronic disease in a region/country can be measured more comprehensively, reporting results that avoid double counts and minimize selection bias in a systematic way. The advantage is evident if researchers are also keen to consider multi-morbidity and the potential of registries from different non communicable diseases.

Population-based diabetes registers aim to overcome the above limitations by collecting data on a well defined denominator: the entire population. However, when the disease has a high prevalence, such a target may be ambitious and will require strong institutional and regulatory support of local governments to be sustainable long term.

The simplest (and faster) solution to create population-based registers is to integrate multiple data sources that are already established and sufficiently complete through linkage of secondary data. There are different advantages and disadvantages of this approach, but in general it sounds attractive for the limited cost compared to primary data collection. There have been many positive experiences of computerized linkage worldwide that have been widely documented. In general, the literature shows that the integration of clinical and administrative data sources to cover entire regions or even countries is easier to realise in health systems providing universal coverage. At the same time, the approach seems more feasible in geographical areas of limited size (sub-national), or within the boundaries of a population of 5 million inhabitants.

3.2. Impact on the calculation of health indicators

The way registries are organized may have a clear impact on the calculation of target health indicators. The nature of this problem is shown in **Figure 3.1.**

Indicators are frequently constructed as rates or percentages of occurrence of a certain condition within a well specified group of subjects. Typical examples are the percentage of patients treated with beta blockers after myocardial infarction, or mortality within 90 days post infarction, or lower extremity amputations in people with diabetes in a specified year.

Although apparently simple, indicators require many complex rules (e.g. inclusion/exclusion criteria) specified at the outset in technical guidelines (e.g. the OECD health care quality indicators, see Carinci et al 2015a) to form numerators and denominators (the "cocktail" in the right side of the Figure). Furthermore, the calculation of standardised rates may be hampered by incomplete data if multivariate risk adjustment is also required.

The availability, contents, and interoperability of data sources may induce bias in the calculation of indicators. For example, in the calculation of mortality rates post-infarction, if the personal identifier is not reliable, then it would be impossible to link hospital admission data to a mortality register and even to readmissions to other hospitals. In diabetes, the type of diabetes may be required to stratify results by relevant subgroups, or a diagnosis date (even the year) may be essential to adjust estimates by duration of diabetes as required by clinicians.

A population-based register can reduce bias in the calculation of indicators. In this case, the reference population is the total population in a specified region (large green circle on the left). Cases are extracted from this population through a centralised list, e.g. a client master index, which can be used to link to any service provided or relevant episode via a Unique Person Identifier (UPI).

In this way, a relational database can be formed linking across data sources, and all numerators (outcomes) and denominators (reference population) are well formed, minimising the likelihood of double counts and mismatches. Longitudinal cohorts and the analysis of clinical pathways is simplified, as well as the application of sound epidemiological methods.

The scenario is completely different whenever a centralised list is not accessible and registries are formed from one or more linked health care databases managed by individual providers e.g. outpatient clinics, specialist services, or primary care centres. This is frequently the case, in which records are created for each single visit.

Under these conditions, several problems arise:

- different measurements are done at different visits, which may imply that values that will be analysed do not refer all to a specified point in time
- the total population is not that of a common region, but is the union of the catchment areas from all providers included in the register (pale green circle)
- the denominator of people with the disease (green circle) may substantially deviate from that of the region because: a) it would not include subjects that have not been seen ever

during a reference timeframe (usually one year) – which may be the sicker ones; b) it would include patients outside the region that lie within the catchment area of one of the providers.

- basically each provider contributing to the register would only include in the denominators the list of "active patients" (yellow circle), which can undermine the reliability of all indicators depending on the extend they deviate from the overall number of people with the disease in the catchment area. Take for instance the "percentage of patients with at least one measurement of HbA1c in one year". If the visits/records are created at annual checks, this indicator is very likely to be significantly biased upwards or even close to 100%. That is because the numerator does not take into account individuals who have not done a visit at all, which may be in turn associated with the outcome of interest. For instance, measuring blindness with visit records may be severely biased, given that those with reduced sight may be less likely to make regular visits and thus could be excluded by the active list.
- as a result, outcomes in a population with a specific disease (alias numerators, white circle) may emerge as a combination of different sources of bias. They are certainly within the catchment area, but can be within or outside the population of the area. Depending on how many active patients providers have involved, they can refer only to a certain percentage of the actual population with the disease. For these reasons, in many cases the results of the outcomes on provider-based registers may be severely biased, thus not representative of the true underlying population. They may be usually optimistic, as the provider has the capacity to manage the list of active patients according to own preferences.

To reduce the impact of the design of provider-based registers on the potential bias, several best practices can be highlighted:

- agree on a standard definition of "active patients" to increase the ability of including those "hard to reach" e.g. "delete only individuals who did not make a visit in three years". An objective method could be defined through targeted sensitivity analyses;
- agree on a standard method to merge different measurements done at different visits;
- agree on a common method to process records for catchment areas extending beyond the reference region;
- include in the register a regular update of population-based contextual data e.g. total population and estimates of the population with the disease (or prevalence) stratified by age, sex and all relevant confounders.

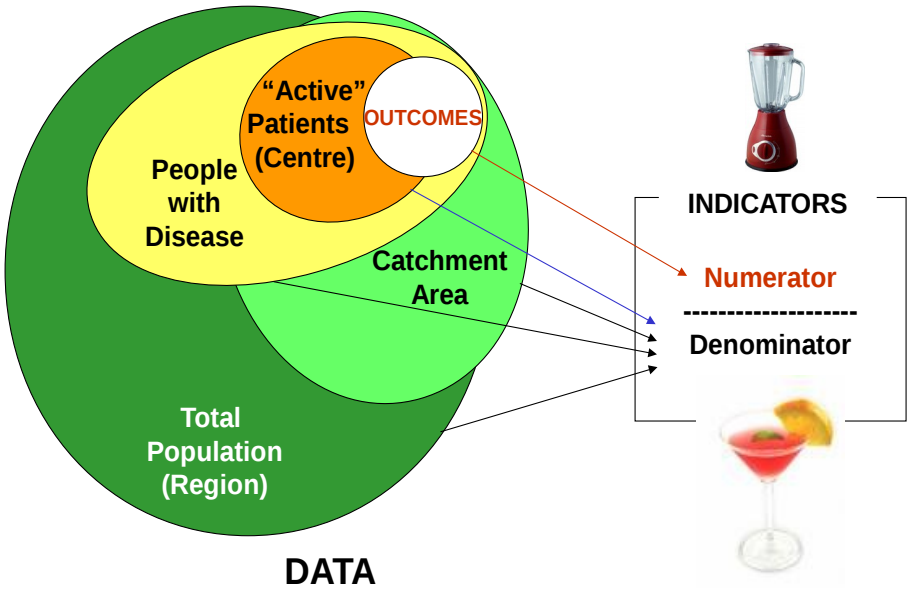

Figure 3.1. Problem of bias in health indicators

CHAPTER 4

The difficult integration of data sources

4.1. The intrinsic heterogeneity of data sources

Structured monitoring systems may be paramount to plan effective strategies to contrast the increase of non communicable diseases through targeted interventions in the areas of care and prevention. However, complete population-based registers may be difficult to implement in the short term for both technical and political reasons. In these situations, it may be advisable to invest on the integration of the existing data sources, which are often heterogeneous in many respects.

Data sources may have been created for different reasons, e.g.:

- to support disease management programs with networks of health professionals
- survey using a representative sample of the total population for statistical reason
- quality benchmarking of health services using linked administrative data
- evaluation of policy and planning in a specific territory

In all the above situations, data integration under a common framework (e.g. national or EU register) may be severely hampered by the heterogeneous contexts in which data are generated. However, using all the existing information would be extremely convenient, as it could help not only policy and planning, but also improving the information infrastructure through continuous use of the available data.

Leaving all legal hurdles of privacy and data protection aside, two solutions seem immediately possible:

- Merging all data into one big centralised database and mapping (conversion) of the original definitions towards commonly agreed rules.
- Imposing new standards on the existing data collection (redefinition of data structure)

In both cases there are evident limitations.

In the first case, the creation of a centralised dataset would have clear advantages for management and governance: the system would be available as a whole for programming, secure protection and standardized mapping from a central location. Sensitivity analyses and tests could be also continuously conducted. However, the complexity and cost of maintenance would be hardly sustainable and the level of participation of original data custodians drastically reduced in comparison to a distributed environment. This would also mean that continuous data quality improvement and progressive standardization at source would be hampered. Moreover, the level of coverage of the central database could be also limited by the willingness to submit original data to the coordinator.

On the other hand, imposing new formats of data collection can be extremely inappropriate both for cultural reasons (local practitioners may be very reluctant to change their standards) and for the additional costs required to change systems and train professionals to collect data in different ways.

As a result, a much better strategy could be that of considering the heterogeneity of local data sources practically inevitable, as a part of the cultural values of the community.

The real challenge is to convert the original formats towards common agreed standard that can be evidence-based. The best participatory strategy could be sharing the review of the evidence as well as the interpretation of the results to the local user, who can be put in charge of mapping data from the original definitions towards the newly agreed definitions.

In summary, these are the steps that could be advised to create a meta-register from the available data sources:

a) review the existing standards used in different sources
b) conduct a review of the evidence from the available literature
c) agree on common standards by comparing the results of the two steps above
d) disseminate agreed definitions
e) ask local data custodians to map their data independently towards agreed standards

More details on this procedure are provided in the presentation of the steps of the BIRO approach.

4.2 Evidence-based data dictionaries

The process explained above can be realised effectively using a specific methodology that adopts an open source logic.

A data dictionary, or metadata repository, can create a "*centralized repository of information about data e.g. meaning, relationships to other data, origin, usage, and format*". Specific examples in the health sector adopting the standard ISO/IEC 11179 include the Australian Metadata Online Registry (MeteOR), the US Health Information Knowledgebase and the National Cancer Data Standards Repository. The EUBIROD project has delivered a standardized EU data dictionary for diabetes using this approach (Cunningham et al 2015). The standard can include an XML representation that can be embedded in the information infrastructure via shared open source software.

Figure 4.1 presents the automation of the entire process. Original data sources are as heterogeneous as different type of fruits. The data dictionary corresponds to the “recipe” to deliver the final “cocktail” of indicators. Through the shared definitions, the redundant heterogeneity of the original data is discarded and only the bulk of the data that complies with comparable definitions is left in the database. This method may impose hard rules on data quality that can lead to the exclusion of large sets of records where, for instance, missing data for mandatory items are found. In diabetes, the unavailability of either type of diabetes, date of diagnosis, or episode date may lead entire data sources to be practically unusable. However, these limitations may considerably improve the ability of the system to process only representative data.

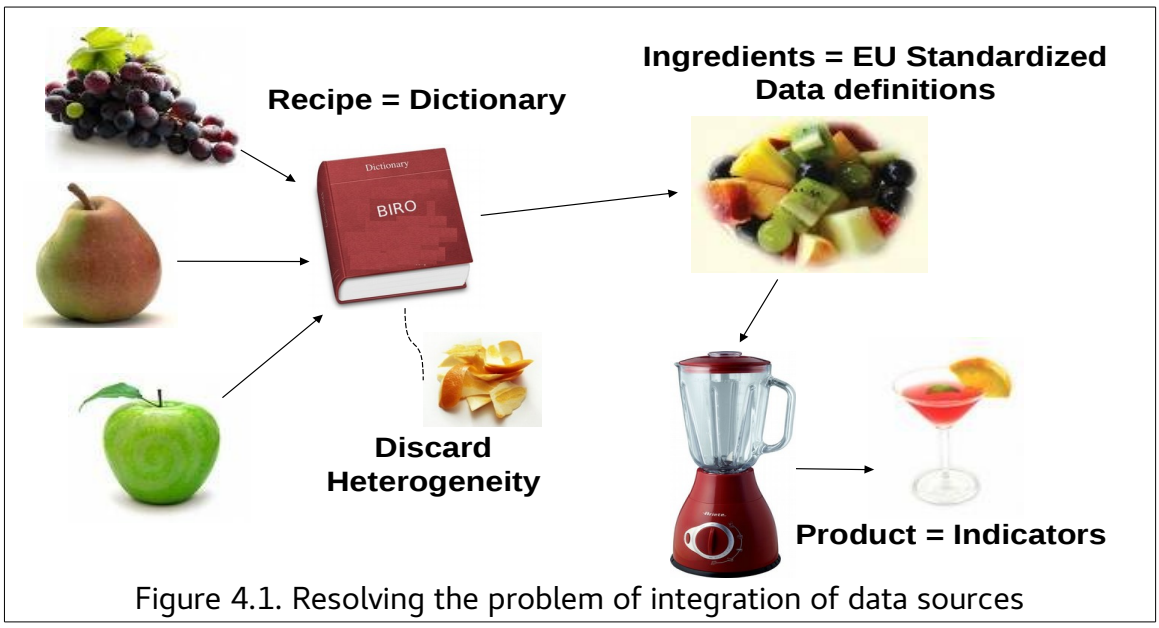

Figure 4.1. Resolving the problem of integration of data sources

CHAPTER 5

5.1. Why do health systems need registries?

A modern disease register represents the fundamental pillar of regional/national strategies against chronic diseases. Through the standardized use of electronic medical records and the adoption of evidence-based definitions, registries can feed a continuous quality improvement cycle for the conduction of first class research, efficient health care management, and quality-controlled routine care for people with diabetes.

However, developing modern electronic diabetes registers requires substantial organizational efforts and pose significant challenges, as they go well beyond the establishment of computerized databases in medical practice. Their role can be central to the automation of a sophisticated network, where all aspects relevant to improving the condition of the individual (structures, processes and outcomes) are constantly monitored through the adoption of common standards and clear targets for the specific population (epidemiological denominator). Such an integrated framework requires linking information from different sources using clearly defined sets of criteria (meta data), e.g. clinical definitions that can quality-assure the content of databases to be exchanged. For instance, the year of diagnosis and type of diabetes can be set as mandatory for each subject included in a diabetes register.

Electronic registries can be used to rapidly obtain epidemiological measures that only few years ago required complex and expensive studies. They can produce independent estimates of standardized rates for a range of indicators and may be used to explore the relationship between potential risk factors and different outcomes of interest, taking properly into account the particular clinical and socio-economic characteristics of the population (case mix).

Across the last ten years, disease registers have been increasingly used:

- to provide robust and timely information on the epidemiology of the disease(s) and associated complications
- to monitor the disease(s) across time, interventions, and changes of the environment
- to evaluate the quality of care delivered to people with the one or more chronic conditions
- to estimate the cost of the disease(s)
- to estimate the cost-effectiveness of interventions
- to provide a solid platform for shared care
- to provide an essential tool for research

Typically, a broad range of users may benefit from the existence of surveillance systems: National policy makers; Health care policy makers; Health care administrators; Health care deliverers; Diabetes research institutions; People affected by diabetes; The public domain.

The well-designed register directly involves all members of the patient's health team, including physicians, nurses, physician assistants, and office managers, and can be used in the process of care, as well as to assess quality of care and health services performance.

The availability of timely information on high-risk sub-populations may allow the healthcare team to better target their care and evaluate adherence to treatment guidelines in relation to the actual trends observed in their routine activity.

Registries have been key components in numerous disease management initiatives addressing data collection at the point of care and systematic information exchange across a network, most often organized at the regional level.

5.2. Case studies in diabetes registries

This section will present cases collected by the Study Group at the 1st EUBIROD General Assembly, sponsored and organized by the University of Surrey on 24-25th August 2015, in conjunction with activities in progress for Bridge Health. Country presentations are an extract of what is available from the proceedings of the meeting.

In **Belgium**, specialized hospital-based multidisciplinary centres are legally obligated to regularly provide extracts for insulin-treated diabetes patients from their medical records, using a standardized electronic questionnaire, for the purpose of quality monitoring and improvement. In addition, there is also a registration among diabetes patients treated in primary care and regular audits in paediatric diabetes centres and specialized centres treating diabetic foot ulcers. The main audit runs every 18 months, with data collected retrospectively on a sample of 10% of the patients. Data linkage is allowed to the extent allowed by the authorisations from the Belgian privacy commission. Collected data are coded, i.e. they pertain to a single theoretically identifiable person, but the record identifier is turned into a code without meaning. Coded information is stored centrally in facilities that conform with data protection legislation. The data custodian is the Healthdata.be department of the Belgian Scientific Institute of Public Health (healthdata@wiv-isp.be). A new integrated platform for data exchange is being constructed by the Institute. Quality of care indicators are developed through e-health solutions capturing data from providers. These activities are funded by the National Institute of Health and Disability Insurance (NIHDI), which pays diabetes centres or general practitioners. Results are fed back to diabetes centres for anonymous performance benchmarking. Collected data are also used for research, exploring different areas, including risk factors e.g. cholesterol levels and diabetic foot monitoring (Doggen et al 2014), with some notable international collaborations with Germany on audit-feedback combinations (Germany).

In **Croatia**, the National Diabetes Registry was established in 2000 with the aim of improving health care of persons with diabetes mellitus, assessing the prevalence and incidence of diabetes mellitus and its acute and chronic complications, monitoring morbidity, mortality and other clinical care quality indicators on a national level. Since 2004, registration has been mandatory for all general practitioners and hospital physicians treating persons with diabetes mellitus. Data collecting in CroDiab is based on BIS (basic information sheet) internationally recognized as the optimal data form for follow-up and improvement of diabetes care. The CroDiab NET system integrates electronic patient

records and is able to generate discharge summaries of the patient and send his/her data to the registry in the same time. CroDiab Web is the system developed for data collection online, specifically developed for the needs of Gps. All the data in the central registry database are merged on the patient level. To ensure the accuracy of data, the national mortality database as well as national physician registry is regularly imported to the registry database. Diabetes quality indicators are created annually with a well defined data dictionary. The CEZIH is a central Croatian IT based primary health care information system collecting patients' data from all the general practice offices, other primary health care units (like paediatrics, gynaecology, dentistry etc.), pharmacy, and laboratories enabling e-transfer of medical data to national insurance company and some public heath institutions including registries. Regular link between CEZIH and CroDiab allows covering persons with diabetes mellitus completely. Data extraction for hospitalizations, sick leaves, treatment and laboratory measurements etc is possible. Quality monitoring is performed through routine data feedback and audit, and annual results available at the website of the Croatian Institute of Public Health. The IT structure of CroDiab ensures data protection on several levels. Data transfer via Internet is protected by means of SSL and 128-bit encryption and user authentication. Access to patient data is regulated and limited only to personal data available from official records. Accessibility of data is enabled only to authorize users. Daily back-up of database includes encryption by a 128-bit key.

In **Cyprus**, the electronic register was developed at the Larnaca clinic in Microsoft Access by the Department of Information and Technology Services of the Ministry of Health. It is based on the BIRO common dataset with some additional data being collected and recorded. Since 2005, four more GP centres with five doctors joined the BIRO project in the areas of Larnaca and Famagusta. At the end of each year, the Department of Information and Technology Sevices of the Ministry of Health processes data collected which is presented to the clinics. Reports are used for own audit and quality improvement. The Ministry of Health has also organized courses for data collection using the system, involving healthcare professionals (GPs and Nurses). The Department of Information & Technology Services of the Ministry is now working to link services from all hospitals towards a centralized database.

In **Denmark**, a National Register covering the entire population of 5.4 million people between 1995-2006 was built through linkage of different national registers (Civil Registration System, National Patient Register, and the National Health Service Register) (Carstensen et al 2011). The register, relying on a reliable personal ID used across the country, allowed to compute fine estimates of age-and sex-specific prevalence, incidence rates, mortality rates and standardized mortality ratios relative to the non-diabetic part of the population. Results showed an increase in prevalence by 6% per year, stable incidence and a decrease of mortality by 4% per year among subjects with diabetes, against 2% per year of the residual portion of the population. The mortality rate decreased 40% during the first 3 years after diabetes diagnosis. However, this platform does not integrate clinical data, and indeed is not capable of discriminating between different types of diabetes.

In **Germany**, several initiatives are aiming to strenghten diabetes monitoring and surveillance through improved data collection (Jecht M et al. 2015). A register for the diabetic foot, a National Initiative called DIVE (Diabetes Versorgungs-Evaluation - National initiative for quality assessment in diabetes care) and a new National System for Diabetes Surveillance. The register for the diabetic foot (http://www.fussnetz-koeln.de) includes an expert group that has determined the parameters and various activities e.g. publications and financial issues. The group is multidisciplinary and data quality is controlled on a continuous basis within a quality of care improvement process. The DIVE initiative (http://www.dive-register.de) includes 199 centres following 135,803 patients. The National system is being developed under the aegis of the German Diabetes Association, the Federal Ministry of Health and is coordinated by the Robert Koch Institut (RKI). The RKI has recently constituted an international scientific board (including members of the EUBIROD network), which gathered for the first time in July 2016 to agree on the foundation principles of the national data collection. In terms of computerized systems, The diabetic foot register operates through a software suite including interconnected tools. DIVE uses as tools DPV and DIAMAX, which are both diabetes management systems to collect and send data to the central server. The infrastructure for the National Surveillance system is still to be determined. Reports are published by all groups to regularly communicate the state of the art to the public.

In **Israel**, the program of National Quality indicators for community healthcare is a joint partnership of the four healthcare plans and an academic research institute. The program collects annual reports extracted from the electronic healthcare record. Reports include a list of quality indicators encompassing various fields in community healthcare. Data is reported as rates and there is no element of follow up. In regards to diabetes the program provides data about prevalence and about the percentage of patients performing required test and that are controlled. The need to have more extensive data about diabetes on the national level including information regarding incidence of diabetes and its complications led to the formation of the registry. The National Diabetes Registry was formed in 2013 with the aim of providing epidemiological data about prevalence and incidence, plus information about the trends of the disease and rate of complications, as well as a basis for research initiatives. The registry is managed by the Ministry of health and operates in full partnership with the four health plans that provide ambulatory medical services to the citizens of Israel. All citizens have national health insurance and are members of one of the health plans so population coverage approaches 100%. Reporting is done annually. Data is extracted from the EHR of the health plan and include demographic details and a few clinical parameters. Identifications of the diabetic population is based on the results of blood tests (HbA1c / glucose) preformed in the previous year or the purchase of anti-diabetic medications. The reports include individualized de-identified data coded by a coding mechanism that enables cross linking with other databases that are coded with the same algorithm. Using the identification of diabetes by the health plan as the standard, the positive predictive value of the register was estimated to be equal to 93%.

In **Latvia**, the Diabetes Register was set up in 1997 and data collection was based on the

form of DIABCARE II. Since 2008, the register is part of a web-based Register of Patients with Specific Diseases, set up and maintained by the Centre for Disease Prevention and Control (CDPC). The electronic version, which includes most of the characteristics listed in the BIRO data dictionary, has the primary aim of developing a unified data information system about patients with specific diseases, ensuring the provision of accurate statistical information across the country. The Register is population-based and collects new diabetes cases as well as updated information for each patient (70-75% of the total prevalent cases) in the Register once a year. The CDPC uses own quality check plan to control data accuracy, user operations and other relevant quality measures. Results are compared to those released by the Latvian NHS data, to inspect about differences in rates e.g. lower extremity amputations. Since 2012, the Register is linked to the Population Registry, so that accurate personal characteristics could be also added to the database. The Register is regulated by a framework of the Cabinet of Ministers Regulation passed in September, 2008. Its implementation and maintenance is funded by the national budget and run by CDPC. Health care institutions do not receive special funding for data provision, mainly carried out by specialists (endocrinologists) and GPs (primary health care specialists). The Register is also linked Death Causes Register, so that a person can be automatically removed from the denominators after death. In terms of IT infrastructure, data entry is done using *ad hoc* software PREDA (Patiens Register Data), which is web-based and uses a secure data transmission channel. The system databese operates on MS SQL and XML WEB service. Information from the Register is used for regular reports, policy framework document planning and evaluation, as well as research (Pildava S et al 2014).

In **Malta**, the Malta Diabetes Database has entries for basic patient demographic data, diabetes and medical history, diabetes and non-diabetes drug entry, physical examination and complications List, dietary history and advice and education. Data is collected from seven diabetes clinics though data collection and is not 100% complete. The system has common fields with the BIRO system and can automatically generate reports for analysis by the BIRO system. The ata custodian is Professor Joseph Azzopardi.

In **Norway**, the Norwegian Diabetes Register for Adults (NDR-A, Cooper JG et al. 2013). was granted status as a consent-based National Quality Register by the Ministry of Health and Care Services and approved by the Data Inspectorate in 2005. The main aim of the register is to improve the quality of treatment for people with diabetes. The register also provides data for research on diabetes and diabetes-related conditions. NOKLUS is responsible for the development and day-to-day running of the register, funded by the Western Norway Regional Health Authority. Haukeland University Hospital owns the Register and is responsible for ensuring that the health data it holds is handled in compliance with the Norwegian regulations that apply for the processing of personal data. NDR-A has an annual budget of approximately NOK 2,4 million. Participation in the NDR-A is not mandatory but all general practitioners, specialists and hospital outpatient departments in Norway are invited to participate. Recording of data requires written informed consent from the patients. Data collection started in 2009 with 3 hospitals and 100 practitioners reporting to the register. In 2014, 32 of approximately 45 hospital

outpatient departments (71%) and 362 of approximately 4000 general practitioners (9%) reported to the register. Current estimates suggest that approximately 25 000 patients have type 1 diabetes and approximately 200 000 have known type 2 diabetes in Norway. The register has data on 34% of the patients with type 1 diabetes (8407 patients) and 8% with type 2 diabetes (16223 patient). Data from the register has been used as the basis for scientific publications and PhD theses. It provides annual quality reports to participating centers and individual doctors and is used in National guidelines for diabetes care. The data collection is annual. The registration of data is carried out electronically by general practitioners, endocrinologists, specialists in internal medicine, nurses or other health care workers with a special interest in diabetes, when patients with diabetes come to regular follow-up appointments. The patients and data Inspectorate have approved data linkage to The Children's Diabetes Register, The Prescription Register, The Kidney Register, the Cardiovascular Register, the Medical Birth Registry, The Cancer Registry, The Registry of Deaths, The Education Registry, The Sick Leave Registry and the Norwegian Patient Register (which includes information about all patients that are treated at hospitals or by practicing specialists). Specific computer software has been developed to improve the quality of data capture and to reduce additional time required for data entry: Noklus Diabetes, which can also import patient identification data, laboratory results and prescription data from the main electronic medical record and also returns a summary note. Noklus Diabetes can also provide decision support and reminders to health care workers. The software is distributed free of charge to participating units. Hospital clinics have to pay a modest annual license fee to cover the cost of support and future development work. An annual quality report where results from the local unit is compared with aggregated data from all participating centers is disseminated to the local unit by email or mail. Annual reports are available from the main website (https://www.kvalitetsregistre.no/resultater/).

In **Poland**, the Silesian register, covering Upper Silesia region in Poland (population 4.6 mln, 12.5% of Poland population), is the biggest regional, university based register of type 1 diabetes among children aged 0-14 yrs. The register was created in 1989, as a part of the EURODIAB project. Other similar registries exist in Poland, including the regions of Lodz, Gdansk, and Białystok. Altogether, these registries cover over 30% of the Polish population. The register is able to document the broadest and most long-dated study of Type 1 incidence in children and adolescents from former socialist countries. Currently, the register is starting its expansion to cover incidence of type II diabetes among adults. A recently established database covers all new cases from Upper Silesia from 2008-2013. The different aspects of diagnosis and treatment are under investigation together with analysis of the financial aspects of type 1 diabetes treatment and its relation to socio-economical factors. The registers works in cooperation with the National Health Service and the Polish Ministry of Health, to establish multi-center working groups including clinicians, epidemiologists, bioinformaticians, and representatives of governmental agencies. The database is updated annually and is based on local software. The activity is used for education and research, with several articles published as peer reviewed papers (Chobot A et al. 2015).

In **Scotland**, the Tayside population-based register has been continuously operational since 1996 (Morris et al. 1997). This initiative, deeply integrated with the local government, supports the provision of clinical practice, research and governance through performance reporting. The registry started as a research project, the Diabetes Audit and Research in Tayside Study (DARTS) (Boyle et al. 2001). The project progressively grew from the original local catchment area involving the Tayside Regional Diabetes Network (TRDN), to the Scottish Care Information Diabetes Collaboration (SCI-DC), which constitutes the infrastructure of what is today the Scottish Diabetes Register. In the register, information is split between the research reports (analysis) and the clinical services (enhanced information systems). The register explored, among others, estimates of prevalence, mortality rates, the efficacy of pharmaceutical treatment, hypoglycaemia, diabetic complications, eye disease, and pharmaceutical costs. It allowed the construction of the first validated, population-derived model for prediction of absolute risk of coronary heart disease in people with type 2 diabetes. These algorithms provide decision support tools for clinicians involved in diabetes treatment and indicate appropriate early action to decrease the risk of adverse outcomes. The experience of Tayside shows that data quality of official statistics may also be increased from integrated information. However, information exchanged over a region will never be 100% accurate, but systems may be improved significantly through their routine use and analysis. The Scottish register was born from bottom up, through the direct participation of clinicians and people with diabetes. The national database is currently updated overnight through a network of operating regional servers, allowing clinicians to benchmark their quality of care on a daily basis and almost in real time. In a system like the one constructed in Tayside, it is possible for a clinician to instantly access measurements for a person with diabetes across different providers and compare the same parameter across the average scored for the reference population. These results can be returned to the individual, closing the loop of quality of care improvement and person empowerment. Along the years, strategies encouraging “clean” clinical recording entry minimized “dirty” data contamination in the Tayside register. In this case, the evolving nature of the process has been intrinsic to the collaborative integration of different sources, for which “*the job of creating adequate databases will never be finished, but striving to create adequate clinical datasets will always be worth doing*” (McAlpine R. 2009).

In **Slovenia**, all children and adolescents with newly diagnosed diabetes are referred to only one central institution: the Department of Pediatric Endocrinology, Diabetes and Metabolism (DPEDM), University Medical Center, University Children's Hospital, Ljubljana. This centre is also responsible for the Slovenian National Registry of Childhood Diabetes (SNRCD) since 1970. This registry is cross-checked annually with the reports on ‘cause-of death’ for this age group. DPEDM was also an active member of the original EUBIROD project with providing data and expertise for pediatric section. The Department of Endocrinology, Diabetes and Metabolic Diseases (DEDMD), University Medical Centre Ljubljana, is responsible for Slovenian National Registry of Adult Diabetes (SNRAD). Data collection for this registry started in 1982. The registry operates using database technology based on state of the art and industry standards such as openEHR, IHE and HL7. DPEDM also participates in international data collection and benchmarking as a Centre of Reference

of the SWEET project. The database is used for policy and planning, as well as for research (Dovc K et al. 2014).

5.3. Challenges of information sharing in chronic diseases

Cultural and technological barriers pose significant challenges for the realization of structured information exchange in health care systems. The scale of the problem of non communicable diseases, with their overall burden for the population and implications on the organization of different health services, represent one of the most problematic areas where comprehensive solutions could be possibly organized.

Today, a primary element that cannot be overlooked in the construction of disease registers is that of individual privacy and data protection. The increasing complexity of the legislation in this field has in fact generated a heterogeneous implementation of fundamental principles, producing in practical situations an imbalance between the right to privacy/data protection and the right to health. A revision of procedures in place in diabetes registers from eighteen practices carried out by the EUBIROD project found a high heterogeneity in the application of criteria related to anonymization, consent, accuracy and access to computerized information (Di Iorio et al. 2013b). Such lack of uniform approaches may generate concern for appropriate safeguard of information in health care, which in turn can translate into potential new impediments to information exchange from the revision of relevant legislation.

While it is fundamental that information systems conform to current privacy legislation to ensure their integrity and safe continuation, it is also important that their value for public health and routine is increasingly recognized, to avoid that excessive restrictions are imposed on information exchange and make sure that consent does not become a problematic issue, influenced by mounting concern on threats for personal privacy.

From a scientific perspective, using a computerized integrated register offers new opportunities for research studies based on gold standard methodology. Sampling plans may be facilitated by the availability of a large or even complete pool of subjects, from which groups of individuals may be enrolled in cohort/observational studies, case-control or randomized controlled trials.

On the other hand, an evidence-based comparison of the effectiveness of the existing registries is hampered by the specificity of their implementation, as it would be difficult to isolate the local conditions from a systematic effect of a particular solution. In fact, it would be impossible to “randomise” aspects related to the structural organization of health systems, each with a unique culture and specific policies that are associated to the average outcomes. Nonetheless, it would be still possible, although still uncommon, to compare different alternatives within similar settings, e.g. routine care agains audit and feedback, computerized reminders, mobile health care, aid tools for self-care, etc.

Practical challenges emerge from the analysis of specific operational contexts. The task of building a structured platform for data sharing represents a naturally evolving process, which must be tailored to the the environment where the system is activated.

Population-based registers, albeit methodologically attractive, may not represent the easiest solution to be implemented at the outset. Computerized data linkage of administrative databases using a unique subject identifier, as well as targeted programs of disease monitoring across multiple service units, or a mix of these approaches, may represent convenient alternatives under most practical conditions. By all means, a well planned information infrastructure shall be based on the adoption of standardized definitions and is key to avoid catastrophic investments on inadequate frameworks. In this context, the increasing restraints on data protection shall be adequately taken into account.

The brief review of international experiences in the field of diabetes clearly shows that the creation of a National infrastructure for non communicable diseases is by far the most ambitious endeavour. Even the most advanced health system struggles to implement large scale data warehouses, not only for technological reasons. A more successful strategy would recommend building upon sophisticated systems implemented in regional areas, to extend the approach of direct interaction between all relevant stakeholders (policy makers, health professionals, researchers and citizens) in a federated fashion.

General recommendations should be taken into account in all the above situations.

A disease register must be flexible enough to accommodate the evolution of needs, knowledge, technologies according to the available resources (economical, cultural, and structural). A modern design implies the adoption of a dynamic structure that can embed different sources of information in harmony with the cultural evolution of its own users. Contributors and stakeholders gradually improve their ability to pose sophisticated questions to a system where the overall level of participation is of crucial importance.

The analysis of user perspectives shall become a fundamental element in planning any register, as in any part of the world the design and evolution of these instruments is heavily influenced by different dimensions of the local culture.

At the level of the health system, it is paramount to verify how health care is organized and actually delivered, considering the future impact of modern technology in daily practice.

Societal factors may considerably interfere with the mplementation and regular automation of a registry. Interoperability of systems, common semantics, communication technology and software engineering, database implementation (with a particular attention to the standardization of classification systems), and, most importantly, the local attitudes towards privacy and security legislation, constitute essential elements that must be cautiously evaluated at the outset.

5.4 Architecture of an integrated register for chronic diseases

Building an integrated register involves different roles and responsibilities and multiple actors that would be called to interact continuously with a common system to accomplish different tasks.

An analysis of specific requirements ensuring that short, medium and long term goals are realized must be performed a priori, taking well into account the vision, mission and goals of the register. Such requirements are fundamental to define an implementation plan in

which all stakeholders participate with a double role of contributor/user, making the information infrastructure sustainable and open to quality improvement.

Stakeholders involved in disease registers are citizens, health professionals, health care organizations, policy makers, a National/regional Ministry of Health, academics/scientists, a Coordinating Centre (internal or external to the Ministry) and international partners.

The architecture of the register derives directly from the efficient organization of all procedures involved in implementing such requirements.

Figure 5.1 shows a possible structure of an integrated population-based disease register, reflecting the roles and responsibilities of all the above categories of users.

A system for structured information exchange should consider data provision from primary care centres, specialist services/outpatient clinics and acute inpatient care. To be complete, the system should link records contributed by those sources on a daily basis to other archives managed by public entities, whose content is highly sensitive for personal privacy, including information on a broad range of personal aspects (residency, socio-economic status, etc). A Coordinating Centre may be useful to support the transformation of individual records into a structured dataset that can be used for research and performance reporting. The dataset can be properly anonymised and made available to trusted parties as a research registry, open to collaboration in relevant international activities e.g. a global NCD monitoring network. Linkage and transformation services may be established to respect principles of personal privacy and data protection.

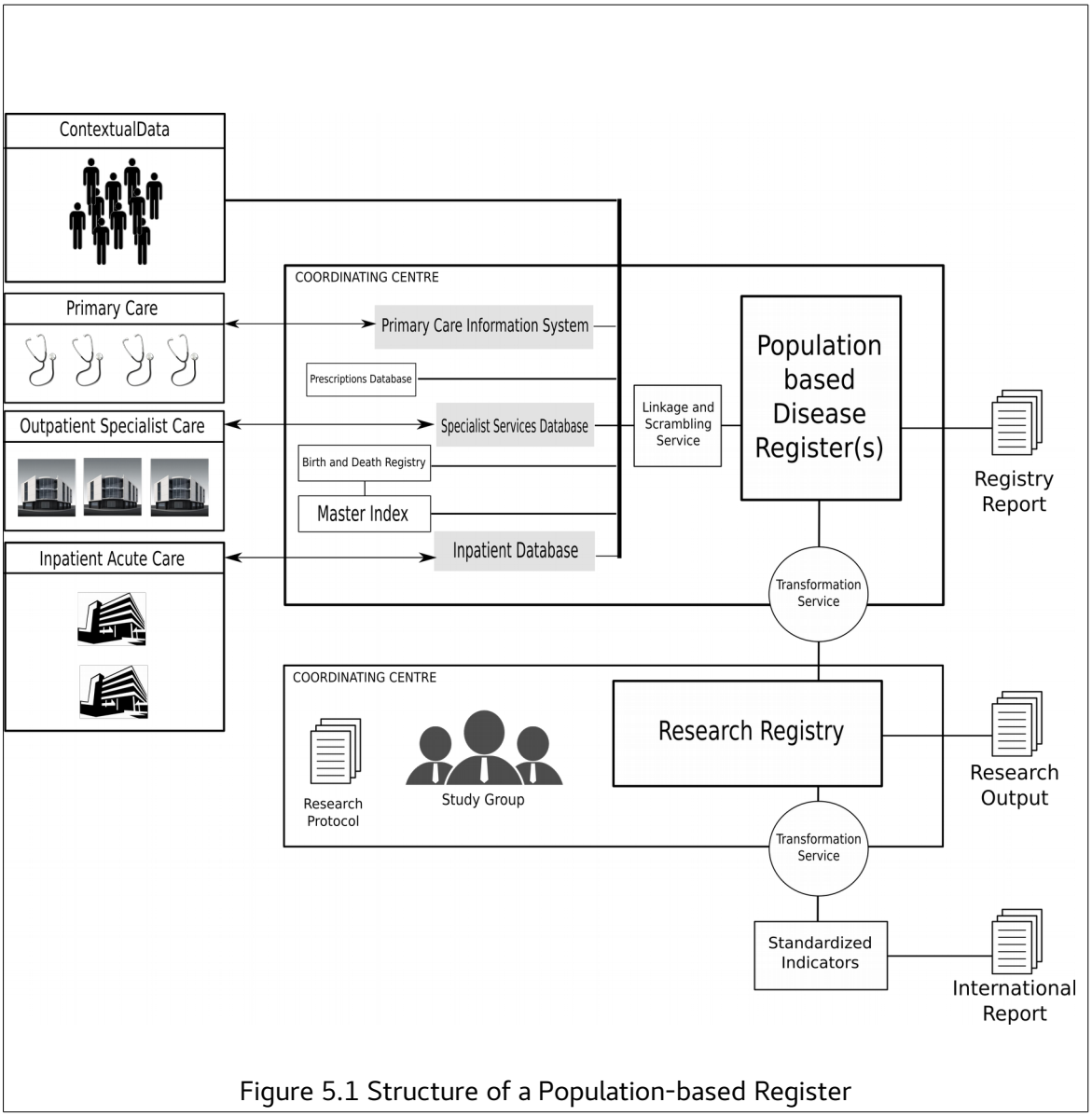

Figure 5.1 Structure of a Population-based Register

CHAPTER 6

Information infrastructure of a EU Population-based Register

6.1 The growing need for best information on chronic diseases

In recent years, international organizations prompted national governments to increase their efforts in activities aimed at fostering quality of care and outcomes monitoring on a global scale.

Prime Ministers and and representatives of national governments gathered at the United Nations 19-20th September 2011 for a High-level Meeting on the prevention and control of Non-Communicable Diseases (NCDs), particularly for developing countries.

The result was a political declaration where Heads of State and Government committed to five key actions: a) reduce risk factors; b) strengthening national policies and health systems; c) international cooperation (including collaborative partnerships); d) research and development; e) monitoring and evaluation. Regarding the latter, the indication was to develop in only one year "*a comprehensive global monitoring framework, including a set of indicators, capable of application across regional and country settings, to monitor trends and to assess progress made in the implementation of national strategies and plans on NCDs*".

A EU reflection process conducted in 2012 resulted into stakeholders providing their advise on how to overcome some of these limitations. In particular, (Van der Wilk et al. 2012), stakeholders suggested some principal changes and improvements in national healthcare systems to better tackle chronic diseases, including "*Improve information and data systems, including disease registries, to better communicate, organise, implement and evaluate the quality, effectiveness and patient-centeredness of national care systems*". The key messages included the following: "a) b*oth at national and at EU level, comparable information, on incidence, prevalence, risk factors and outcomes of chronic diseases is currently lacking; b) Making use of existing structures and activities for data collection is important; c) Information Technology (IT) is especially useful for the delivery of individual patient care (eHealth, tele-monitoring, patient information); d) IT can also play an important role in data provision, e.g. through using IT in the development of patient registries; e) Europe needs mechanisms for safeguarding, providing and strengthening data linkage in the face of privacy and data protection concerns; f) to obtain comparable information common health data collection methods across Europe and permanent co-ordination is needed.; g) To make sure that data and information can actually be used by policy makers and by experts, not only a good data infrastructure is important, but high quality dissemination and reporting activities as well.*

In a following project on "Integrated surveillance of Noncommunicable Diseases (iNCD)", WHO Europe tracked progress in the prevention and control of major NCDs - cardiovascular disease, cancer, chronic lung diseases and diabetes - and their key risk factors. The project reviewed key international databases and the degree to which they already hold data against the indicators of the monitoring frameworks, assessing their completeness and quality.

The results (WHO 2015) showed that various systems co-exist in Europe, which "*make use*

of various types of data sources, from the more traditional population-based registries to health surveys on disease and risk factors, to sources that monitor the content or marketing of food products, according to the population group targeted (for example, children or adults). The systems also differ in approach, for example, from using electronic health records (EHR) or integrating different sources of indicator dimensions to using a multipurpose behavioural community surveillance system. The results of the review indicate that there is a plethora of good practices that could be considered for adoption or adaptation by the countries to strengthen their current NCD surveillance systems, although acceptability, feasibility and cost would be important determinants. The sharing of experiences among neighbouring countries or regions may become an important catalyser".

As a further development, during the last two years the collaboration between the OECD, WHO Europe and the European Union has been intensified through a joint data collection. However, it is clear that huge gaps to make an everyday use of NCD information for policy still remain and would be difficult to overcome if a solution for common data standardization and integration is not found. A bottom up initiative, though, has never been seriously resourced or even engineered. This is where the innovative concept of "essential levels of health information" may turn to be useful and sustainable.

6.2 Pioneering work in diabetes

Consistently with developments in the broader area of NCDs, policy activities were also conducted in the field of diabetes.

A Resolution on Diabetes (61/225) was passed by the United Nations (UN), while the European Union passed a Resolution on Diabetes (14 March 2012) specifying the key action points immediately required to revert the spread of the epidemics. It also included specific mention on monitoring, where the European Union "*Calls on the Commission to draw up common, standardised criteria and methods for data collection on diabetes, and, in collaboration with the Member States, to coordinate, collect, register, monitor and manage comprehensive epidemiological data on diabetes, and economic data on the direct and indirect costs of diabetes prevention and management*".

The EU Resolution was actually preceded by pioneering initiatives funded by the European Commission specifically to resolve the problem. Two projects, EUDIP and EUCID, defined EU diabetes indicators in the early 2000s. A following one, "Best Information through Regional Outcomes" (BIRO), 2005-2009, worked out a solution to publish international reports automatically and on a regular basis.

The basic principle of the BIRO system was that diabetes information already existing in a fairly standardized needed could be rapidly integrated using open standards and privacy-enhanced exchange of aggregate data. In 2009, the project delivered a prototype that allowed collecting seventy-nine indicators from a network of diabetes registers, publishing an international diabetes report in almost real time.

The sequel project "European Best Information through Regional Outcomes in Diabetes" (EUBIROD) aimed "at establishing a European Diabetes Register through the extension of

the BIRO network and the use of related technology" . Completed in March 2012, the system delivered fully versatile BIRO software that has been successfully used to collect data and deliver an international diabetes report from twenty-one countries.

The advantages of this approach under different conditions, e.g. linking data from developed and developing countries, were entirely evident. The BIRO system used a distributed approach, avoiding the storage and processing of huge amounts of data at a server location. By doing so, it can foster a bottom-up approach, preserving the original data ownership, and guaranteeing more strict data quality control through the direct engagement of the data custodian. Sound statistical procedures including risk-adjustment methods were applied to help tackling selection bias and delivering standardized comparisons at EU level. By the way, open source software can be available at no cost but require proper training. For this reason, dissemination activities e.g. the "BIRO Academy" were also conducted to support a cohesive program.

The future legislative trends in the area of integrated diabetes registries make implementations e.g. EUBIROD even more strictly urgent. Following the release of the new Global Directive on Data Protection, "privacy by design" is no more an option, and this was implemented at the outset by EUBIROD.

The EU Directive on the application of patients' rights in cross-border healthcare [23] promotes collaboration between Member States and exchange of information to enable continuity of care and patient safety across borders. According to the Directive, by 2014, a person with a chronic disease should be able: a) to access information on the average quality of care provided by accredited centres in Europe; b) to bring personal data to a provider located in a different country, and be able to add own data to the local register; and c) to extract personal data from the local registry and be able to transfer its content back to the country of origin. On the other hand, care providers should also be able to reciprocally exchange personal data, according to the needs of the patient. These goals can only be achieved by standardization of methods and tools used in different countries. The availability of common guidelines, methodology, and standards to harmonize national registries is the main theme of the EU-funded Joint Action "Cross-Border Patient Registries Initiative" (PARENT) started in 2012. Another joint action, JA-CHRODIS is also tackling the issue of chronic diseases from a broader perspective.

The current trends show that there is still significant work required to transform the positive initiatives undertaken in Europe into a permanent system for monitoring and surveillance of NCDs.

6.3 Essential levels of health information for chronic diseases

The idea of "*essential levels of health information*" (Carinci 2015b) may be useful to finely tune the products envisaged by Task 8.2 of the Bridge Health project. This section draws upon the contents of the above paper, adapted to the specific needs of NCD surveillance. The production of health information spans across a continuum linking input data elements to the outputs that shall be communicated to the public. In between, there are all processes required to transform raw data into useful information. The most natural and efficient way

to identify a common framework is through the integration of approaches from successful EU projects conducted on field over the last 20 years. The task of identifying a common model for different cases may be particularly challenging.

An example from the end of data inputs is given by the case of multiple chronic conditions. Although keeping indicators for each individual disease would be still relevant, new composite indicators are needed to investigate the compound condition. Their calculation will require either ad hoc surveys or linking data for the same subject across multiple sources. The latter case would allow automated calculation, but require disease registers that are structurally interconnected. Such an approach, partially experimented by the EUROHOPE project (Hakkinen et al 2013) in five countries (Finland, Hungary, the Netherlands, Norway and Sweden), seems almost impossible to realize widely across Europe.

One aspect that raises particular concerns is compliance with EU privacy legislation. The impact of the new Global Directive for the practical needs of research is still in its early phase and will need ad adequate time to be fully understood.

At the opposite side of the process, a common model for the communication of the results would also require integrating views and perspectives of different categories of users. Starting from EU projects on chronic diseases may only partially help in this regard, as they have been conceived primarily by researchers, whose views and priorities are very specific compared to others e.g. health professionals and the public.

Recent experiences in the public provision of health information show that targeted efforts are needed to make dissemination more effective. A multidisciplinary approach is required to entail scientific integrity into a broader communication strategy that would allow for results to be used by different categories of users. This aspect cannot be overlooked, as health professionals and the public have their right to shape future health information for Europe.

Therefore, both the design of data specifications as well as their use by different users must be formally evaluated using objective criteria.

A coherent and sustainable infrastructure that can adequately represent different needs and priorities is outlined in **Figure 6.1**.

The diagram envisages a set of interrelated components, classified as follows:

- “content domains”: general principles of policy goals, ethical and social values (represented with black rectangles); subjects of public health monitoring and health systems organization (grey rectangles); statistical methods and information technology (white rounded rectangles)
- “action levels”: four vertical dimensions, including: disease-oriented clinical definitions, population-based disease register in a local setting, information exchange and the meta-registry platform.

This way, any change may trigger adjustments in one or more aspects of the overarching framework. Arrows included in the diagram show the main interconnections occurring between the different components of the system.

The scheme can be particularly useful to fulfil Directive 24 on the cross-border provision of health services. For its implementation, the routine use of health information will require standardized EU definitions, structurally linked to clinical guidelines and underlying data elements, to be routinely applied to national systems of data collection, in compliance with privacy and data protection laws.

The framework of a EU population-based register would work at four different levels:

Level I. Disease-oriented clinical definitions (local standards). The clinical problem should be well specified at the outset. An evidence review would represent the best starting point. From a EU perspective, it might be useful to refer to relevant public health projects conducted and official EC reports targeting NCD surveillance. Target indicators should be targeted according to sound scientific criteria. This process would allow defining the initial content for the EU Population-based register, including standardized definitions and technical specifications for all indicators, as well as specific methods for their analysis. A relevant example is offered by the OECD HCQI Technical Guidelines for Data Collection: every two years, a set of manuals are distributed to Member States, including basic references for indicators by section (e.g. acute care, primary care, etc), specific criteria to map different data sources, items and classification schemes towards common definitions, and algorithms, including flow charts and software distributed for quality control and on site statistical analysis. Specific tools shall be planned and duly implemented by Eurostat to aid database design at all levels. It will be possible to optimize the approach by defining ELHI that will allow using the same structure of the database for both national and EU analyses. A rigorous taxonomy of all data elements involved in the automatic calculation of target indicators would be also required. So far, duplicated efforts have left room for inconsistent data (e.g. different blood pressure ranges in cardiology and diabetes), resulting in measures that are either misaligned, or lack any mechanism to ensure regular update. A structured inventory of these elements ("data dictionary") is needed to set common terms of reference for the EU-HII. Relevant EU projects shall be also screened and systematically assessed, making sure that indicators can be produced consistently and reliably over time. A repository of standards (meta-registry) would be useful as a means to store information on all relevant data sources, specifying their content, and how the data elements included in the dictionary can be extracted at all levels. The repository might also include a library of open source software for interoperable analysis, e.g. mapping tools to link definitions across different classification systems (e.g. ICD9 vs ICD10) or statistical software for computerized reporting.

Level II. Local Population-based register. Here the essential levels of health information would be practically implemented for the creation of a register for NCDs. This would mean not only providing the data elements required to fulfil EU obligations, but also maintaining a surveillance system that regularly uses health information to report on NCDs locally in a standardized way. The quality and completeness of available information can be substantially strengthened through the active use and feedback from local users. To share common principles, it will be essential that each register specifies how public health challenges would be addressed e.g. diseases and risk factors at high prevalence; multi-morbidity and ageing; integrated care; pay for performance schemes, systematic evaluation

of structures, processes and outcomes; patient reported outcome measures; costs; equity. Data linkage across multiple sources should be also made possible. This would trigger the use of advanced statistical methods e.g. risk adjustment, and application of computer programs developed in accordance with agreed requirements. Here a a particular attention shall be given to compliance with privacy legislation. All data and analytical processes shall be subject to a process of privacy impact and performance assessment, envisaging changes in the local implementation e.g. risk mitigation strategies (including distributed databases, fragmented data analysis, etc). The overall approach is summarized by the principle endorsed by the EC of "privacy by design", according to which privacy and data protection should be integrated into the design of Information and Communication Technologies, which should not only maintain security, but be designed and constructed in a way to avoid or minimize the amount of personal data processes (European Commission 2009).

Level III. Information exchange. Precise specifications shall be provided on how to compile national results (macro and/or micro and/or meso aggregates), and which formats and security rules (e.g. encryption) shall be used for cross-border data transmission. This step will require stewardship of the EC (particularly Eurostat) in consultation with MS and data protection authorities. A key regulatory aspect of data transmission and storage would be data ownership, in agreement with the rights of different actors that will be directly or indirectly contributing to the EU Population-based register. Transmitted data that are de-identified may still belong to a person, group of individuals, care provider, region or country. Health information can be processed and published in various ways, raising ethical issues on the rights of groups of individuals, professionals or citizens of entire territories, who can legitimately request control over own records. The issue on who can legitimately claim rights over the data sent to the EU must be also clarified. This topic often neglected should be carefully discussed and clearly resolved in the interest of governmental institutions, data custodians and individuals.

Level IV. Meta-registry Platform At this level, results from MS will need to be pooled, analysed and publicly reported. To make this possible, a central database shall be organized, including all data elements required to compute EU indicators, through the application of advanced statistical routines. The structure of the central database will strictly depend from the formats agreed for cross-border transmission, which in turn will determine the specific statistical techniques that will be used to deliver standardized estimates, e.g. using subject-level data in mixed models, or aggregating cell counts to apply logistic regression, or pooling model coefficients in a meta-analytical fashion. The central system shall also include effective ways for public reporting, as well as means to provide data feedback to MS (e.g. raw data to allow benchmarking national/sub-national results against international averages). To identify the best solutions, the EC shall take advantage from focus groups included in the governance structure above specified at EU level. The platform should be submitted to an independent ethical assessment and overall evaluation, whose results may trigger a revision of parts or even the entire framework.

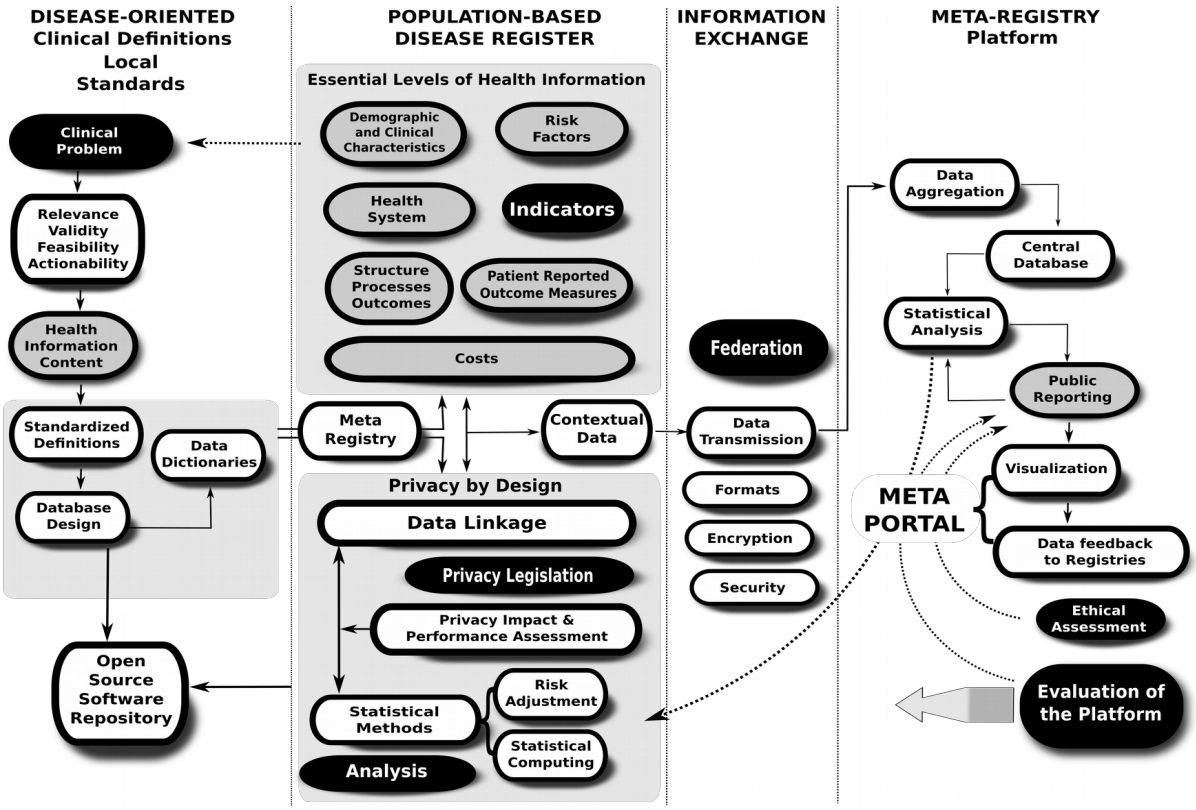

Figure 6.1. Essential levels of health information for chronic diseases

CHAPTER 7

Extending the BIRO approach to other chronic disease conditions

7.1 The clinical-epidemiological perspective

At the 1st Bridge Health Meeting of the EUBIROD Network (Carinci F et al 2015c), the Task Study Group agreed general guidelines on how to extend the BIRO approach to other chronic disease conditions.

The following aspects were highlighted from a clinical-epidemiological perspective:

- Clinicians and caregivers involved in the provision of care for chronic diseases require information:
  a) to make decisions
  b) to know expected outcomes
  c) to inform patients
  d) to evaluate performance
- Clinicians need 'comparable data' rather than numbers. In this perspective, task 8.2 should strive to define a robust European data dictionary where proper clinical definitions can be used to analyse outcomes across a range of diseases and provide the core elements for data collection from existing electronic health records.
- The EUBIROD core standards published in a recent paper (Cunningham et al 2015) respond to these criteria strictly in the field of diabetes – which can be relevant as a general method - but it is also important that these kind of documents can drill down to specific complications. For instance, in diabetes it can be important for clinicians to screen not only retinopathy, but inspect differences between subgroups with background/proliferative retinopathy. Scientific papers may not always be the best means to disseminate such knowledge.
- The task of extending BIRO to other diseases should explain how to merge individual characteristics with relevant clinical information, i.e. age and duration of the disease plus metabolic control, type of therapy and presence of other comorbidities. Increased depth may allow generating outputs of utmost interest eg risk tables similar to the UKPDS risk engine, including cardiac risk assessment, etc. These are outputs that are increasingly requested not only by clinicians, but also by individuals affected by a range of chronic diseases and risk factors.

The Study Group also discussed the current strengths to be exploited and weaknesses to be overcome in the BIRO system:

- the existing report should revisit how indicators are presented and organized through a different elaboration of the associated template
- the BIRO reports are in general very detailed and stratified according to type, gender, age groups, duration of the diseases, with results presented through a range of

outputs including tables, box plots, Trellis Bar plots, and box and whisker plots. While these are all desirable features, it would seem appropriate to allow for the user to build own templates. This would be possible with a modular approach to the statistical routines, so that the assemblage of the report may be customised according to the user needs.

- epidemiological outputs may be more finely tuned. In some cases, these could be expanded to cover different ways of comparing groups (eg displaying and comparing means for continuous outcomes, chi-square tests for categorical data, etc). In some cases, too many cells have zero observations and there are just too many comparisons tested to provide summary results in a usable manner
- comparisons between countries and risk-adjusted indicators are available and can be very useful

These elements have been duly taken into account to restructure the previous prototype in a way that could allocate some of the main requests made by clinicians and epidemiologists.

7.2 The person and policy maker perspectives

In the 2nd Bridge Health Meeting of the EUBIROD Network, the Study Group will attempt to integrate the clinical-epidemiological views discussed at the 1st Meeting, with the perspectives of the person and policy maker, taking into account aspects e.g.:

- the person: what can a person gain from information sharing in chronic diseases and the availability of routine reporting
- the policy maker: measuring outcomes for value-based analysis through patient reported measures and costs for the entire care cycle
- feedback for policy and self-care: how to engage stakeholders from both ends?

Representatives of major patient associations and policy makers at regional, national and European level will be invited for the scope and their views compared in a public discussion.

The dialogues will be based upon:

- major advancements in the area of value-based health care, patient reported measures and cost analysis
- results of the OECD Policy Forum “People at the Centre”, to be held in Paris on the 16th January 2017
- review of the general BIRO approach as a means to organize information on different chronic diseases (see following section).

7.3 Outlining a general approach

A platform of open source software for population-based registries is only a means to an end that can be realised according to a structured protocol agreed by all interested parties.

Within BIRO and EUBIROD, the two Consortia built a system through a workplan that served to the scope strictly in the field of diabetes. The scope of Task 8.2 of Bridge Health

is to bring the framework forward to become a general approach for health information in the field of chronic diseases.

The general BIRO approach derived from the original experience (BIRO Consortium 2009) relies on the conduction of 10 concatenated steps, as shown in **Figure 7.1:**

1. **Review the problem: construct an evidence-based framework.** At first, a conceptual framework should be specified, including the dimensions (e.g. quality of care, equity, etc) and their interrelation, which ultimately lead to major endpoints (e.g. health status) in a system perspective. This step leads to the definition of the "essential levels" of health information for the problem of interest (Carinci 2015). Once a common framework is agreed, an updated clinical review should clarify which indicators should be targeted to populate the matrix, based upon an assessment of the criteria of: (i) relevance; (ii) international comparability, particularly where heterogeneous coding practices might induce bias; (iii) feasibility, when the number of countries able to report are limited and the added value can not justify sustained effort; and (iv) actionability, i.e. the capacity that the indicator can be used to change practices (Carinci et al 2015a). These steps allow embedding an evidence-based approach in the definition of an information system for chronic diseases.

2. **Describe the data structure of your network.** Very rarely one or more data sources would entirely conform to a specific standard (unless they are used as a new standard). Therefore, to finely tune the results of the evidence review to what is available on the ground, it is essential that the data collected at each site/region/country is carefully examined and classified using a common protocol. Comparing these results to the ideal specifications proposed by the evidence review may lead to immediate changes in the definition of target indicators, resulting into a more robust environment for data analysis. This is particularly crucial to understand the impact of mandatory characteristics on the retention and use of individual records for reporting: if the system should discard all records presenting invalid data for one or more columns that are not frequently reported in many sites, the overall results could be heavily biased.

3. **Agree on reporting targets: specify report templates.** A platform is useful to the extent it can report data that make sense to the user, leading to actions e.g. decisions for policy, selection of a provider, evaluation for quality improvement, etc. A specific step of the BIRO approach envisages to conduct an activity that assesses the existing practices of data reporting for a specific problem, and identifies the best options available to communicate results through a common template, according to the advise collected by representatives of the community of users. The source code should be specifically designed to populate the template with statistical results.

4. **Conduct a Privacy Impact Assessment.** The principle of "privacy by design" represents an overarching goal that should be realised in a very pragmatic way. Despite of international regulations e.g. the EU General Data Protection Regulation (European

Commission 2016) and the OECD guidelines on Privacy, national (and in decentralised systems, regional) legislation may allow very different practices on data exchange across systems. The design of the platform cannot be engineered from an ivory tower, but should take into account the difference between codes of practice, local guidelines, and in some cases even the perception of those in charge of collecting and storing data at each source. The Privacy Impact Assessment conducted in BIRO is a comprehensive method using a description of data flow, followed by a revised Delphi panel, to rank alternatives with a different degree of compliance with legislation and risks for privacy. This method is very general and can be replicated elsewhere even with further customisation. In EUBIROD, the methodology has been extended to allow creating measurement scales that can be used to benchmark partners in terms of compliance with privacy principles. Bridge Health attempts to link the two methods in a unique approach to screen partners of the EUBIROD, ECHO and EUROHOPE Networks. A targeted instrument, the Privacy and Ethics Impact and Performance Assessment (PEIPA) questionnaire, will be tested for the scope, so that it can be used to apply the BIRO approach more widely.

5. **Identify the best information system architecture.** The best information system architecture is the one that minimises the risk for privacy, maximising the information content represented by the outputs produced by the platform. Identifying such architecture will never represent an entirely objective process, but the Privacy Impact Assessment can help sharing the design among a community of experts. At this step, a panel of experts may gather and compare their views on how the different sets of criteria would be fulfilled at each level of the information production chain. Basic decisions may involve to what extent individual records are processed and linked across different sources, the minimum granularity allowed e.g. minimum cell size for aggregate data, how data is exchanged and transmitted towards the central server, and stored for further reuse.

6. **Specify your data dictionary.** All data elements required to compute each agreed indicators should be duly described in terms of standardized measures and formats allowed, eventually specifying algorithms to map elements across different measurement units. The result will be a complete data dictionary that can be used to automate the creation of a standardized databases through ad hoc open source software (Cunningham et al. 2015).

7. **Design and implement all software.** After the first six steps, all fundamental elements related to the design of the system would be clarified and the source code can be developed accordingly. A series of choices must be made to match software development with the best architecture identified. Major decisions are expected in terms of data flow and data processing, user interface, database and statistical routines, transmission, server storage and user identification, reporting tools etc. Another level of engineering relates to the choice of programming languages and packages to be adopted, with a preference to those (e.g. Java derived) that are more often reciprocally

interconnected and easier to program in a networking mode. A crucial decision relates to the "weight" of the packages and the number of dependencies involved in the different functionalities system (i.e. when a package requires another one to install and operate properly). For instance, sophisticated statistical routines may be run in R, but if these are not needed according to the reporting needs, other options may be preferred. The same applies to database management systems (e.g. Postgres vs H2, etc). Another major decision relates whether and how the system should be multiplatform (running on Windows, Linux and Apple). One option could be to adopt Virtualization (using tools e.g. VM-Ware or VirtualBox), which can be very effective but is also burdensome from both the programmer perspective (need to release an entire operating system or appliance) or the user (reserve resources for running a different operating system, which can also expose to security risks) Otherwise, Java Virtual Machines might represent a viable option which is quite invisible to the user (simple install of a set of Java tools on each operating system).

8. **Analyse data and disseminate results.** Using the platform to analyse data represents a critical step in the production cycle of a platform that should be amenable for routine use. There is a need to coordinate the network very sharply during a series of tests on field that can undermine he credibility of the system, particularly in the initial phase. A development team and group of site testers should be formed for the scope and a means to connect them via Wiki, Slack or Forum agreed. In this phase, it is paramount to listen directly from users, and be prepared to debug and eventually restructure the code if needed. The development team should be prepared to change programming languages and tools, as a last resource, if the practical experience does not support viability of the prototype either in terms of usability or performance. During this step, tools e.g. Team Viewer to follow directly the analysis on site might be useful on a restricted number of partner sites. Members of the development team should be prepared to work a "hotline" for long hours, considering that the early versions may well lead to very long processing times which could be later optimised. At a mature point of development, it might be very effective to organize a technical workshop (or "hackathon") gathering all site testers to test all the various features and highlight bugs and ways to overcome them. New releases should overcome problems identified, until a major release can be extended to all partners and results could start to be examined. At this point, a reserved area of a website could be used to share initial results with external experts. Once the outputs are stabilised and agreed among members of the network, a dissemination strategy could be agreed, including the release of a portal that could be modelled on the reporting template previously identified. Although it is possible to create websites accessible via credentials for registered users, it would be highly preferable both for transparency and impact on quality improvement, that the portal could be public.

9. **Transfer technology.** Not all partners would be involved in the production or test of the system from its start. This can be due to various reasons: either because they are not technically capable, or not willing to, or in many cases not ready to provide data yet,

which is most often the main reason for which a platform to share information is never built. Indeed, in many cases the emphasis of information systems on the point of care is on the data collection process, rather than "data intelligence". That is not infrequently a major cause of frustration and disillusionment of health professionals towards information systems. Open source software frequently meets genuine and passionate collaboration by end users, as it creates a spirit of collegiality towards the development of a pioneering tool. Second users may not be so easy to attract, but could be willing to contribute if a process of technology transfer is enacted to share the know-how behind the production of the platform. In the BIRO approach, we suggest to undertake this step as a separate activity in which each site is given responsibility to lead an area of technology transfer (IT, clinical, statistical,etc) so that the process can be engaging and trustworthy.

10. **Evaluate, improve and update.** Finally, the system should be independently evaluated by external experts, with a report submitted to the development team and all coordinators of the various steps, so that the platform can be improved updated in a continuous cycle of quality improvement.

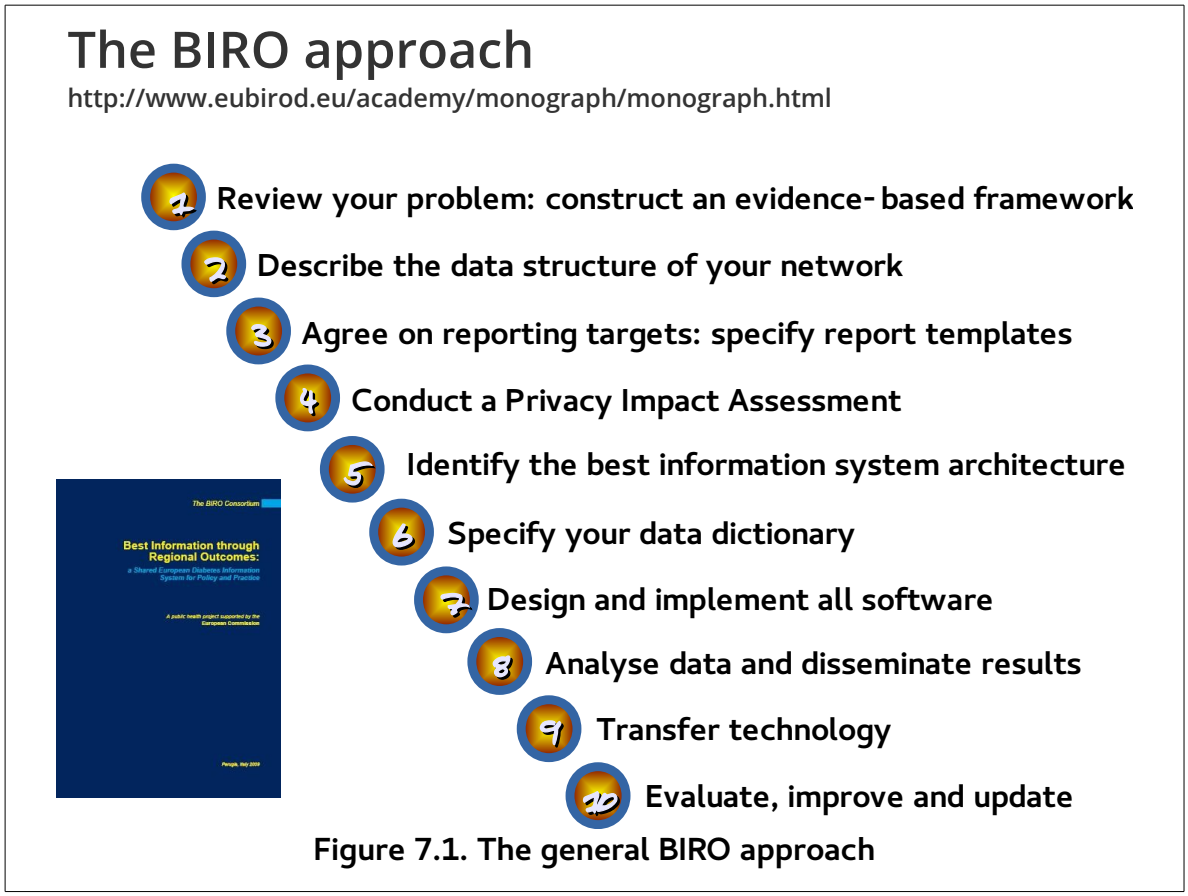

Figure 7.1. The general BIRO approach

CHAPTER 8

Open source software platform for population-based chronic disease registries

8.1 Introducing "NeuBIRO"

Building a common open source platform for population-based registers of chronic diseases requires a comprehensive strategy to resolve complex issues in terms of epidemiological methods, statistical techniques and advanced computer programming.

The platform should allow monitoring the target population across the care continuum, covering all dimensions of the Triple Aim (Berwick et al 2008):

a) health and outcomes of people affected by different chronic diseases
b) quality of care
c) cost of services utilized across the entire care cycle

The system should be capable of processing all relevant data through a user friendly interface that can be used by data custodians without adding substantial work overload to their routine practice, delivering actionable information to a range of stakeholders, including decision makers, health professionals and all citizens (Carinci F et al. 2006). Once widely available, such application can be invoked to perform automated benchmarking across settings, regions and separate jurisdictions in a secure and carefully controlled environment.

The issue of data sharing represents a major hurdle in the health sector. Running advanced analytical routines requires direct access to individual records and agreed mechanisms to interconnect different sources. A federated scheme might represent the best solution to pool databases under a common umbrella.

To be credible, the features of federated databases shall be agreed by a community of users, large enough to be representative of a specific care problem for policy making. Such a system can only be sustainable if the solution is workable and convenient not only as a network, but even for local practice. Therefore, software should be highly customisable. Open source may be ideal under these terms, as the code is directly accessible to the programmer, as well as the algorithms and the annexed know how. Specific licenses e.g. the EUPL* are available to permit unrestricted use and distribution.

The Bridge Health project specifically endorses the BIRO approach (BIRO Consortium 2009) using open source software for the realisation of such a prototype.

The general guidelines for developing the prototype were agreed in the initial phase of the project as a key component of Task 8.2. At the 1st Bridge Health Meeting of the EUBIROD Network (Carinci F et al 2015c), the Task Study Group agreed a workplan for the first 18 months, including the following elements:

- baseline source code of choice
- characteristics of the user interface
- import routines and database engine
- statistical engines

- standard reporting structure
- operational definitions ("algorithms") to produce target indicators

In particular, the group agreed that the development of the prototype should use as a baseline two source code libraries already released by members of the EUBIROD network under the EUPL license:

- the BIRO system realised in BIRO and EUBIROD
- the NEO statistical software (CarinciF, Gualdi S) adapting the BIRO approach to the needs of the Italian Matrice Project (Gini et al 2016)

For this reason, it was decided to name the new prototype as “NeuBIRO”, which will be released along with the present document as a “blueprint” of Task 8.2 of Bridge Health.

NeuBIRO will provide the European Commission with a basic tool that can perform the task of loading, exchanging and analysing data from multiple registers in a very flexible way.

In the following sections of this chapter, we will provide the details of the main components involved in the production of NeuBIRO. The last section will focus on future developments foreseen for the last 12 months of operations. The final document will be available at the end of the project, reporting the overall results obtained for all steps described here. Details on the use and customisation of NeuBIRO are separately described in the annexed guides that will be produced as a separate deliverable of Task 8.2.

8.2 Methodology

The architecture of NeuBIRO is based on the foundations of the BIRO approach described in the previous chapters. Compared to previous releases, NeuBIRO needed to be drastically simplified, overcoming the bottlenecks that were experienced by using the previous platform under real life conditions in the field of diabetes.

General guidelines for user requirements were agreed at the 1st Bridge Health Meeting of the EUBIROD Network,

As far as the production of the **Blueprint** was concerned, it was decided that the main advancements required in the development of the core functions of NeuBIRO were:

- to make it faster, more user friendly, extensible and simple to program
- to minimize dependencies (limited need to install other software)
- to be multiplatform, installed in each operating system (no virtualization)
- to present a configurable input data stream (general across diseases)
- to agree on the format of statistical data outputs to be transmitted to the server
- to include simple routines for data quality control
- to include a simple tool to convert the XML ('Core Standards') to an import specification file
- to foresee link to external data entry software
- to consider data linkage and transformation out of scope
- to allow both local and central analysis
- to be multilanguage, with internationalization module built in
- to embed a simple data transmission module

- to specify client-server communication protocols (eg FTP, SSH, etc)
- to agree upon the organization and governance of the central server
- to conduct a test data collection using only a minimal number of indicators (see above)

Regarding the **Manual of technical specifications for users and programmers**, it was agreed that this should include details on:

- how to use the software web and pdf format
- data requirements and preparation (ETL), e.g. information on the various tables to be used (merge table, population, activity, etc).
- quality issues (clarifications of restrictions for data use)
- legal issues for data privacy (including assessment criteria from the questionnaire released for Bridge Health), explained in short

Among the creative ideas worth further attention it was also mentioned:

- to provide indications on how to structurally link this information to the meta-registry (sort of "BIRO-tunes" e.g. an indicator web repository of draft, approved and domain specific indicators)
- to consider the possible implementation of a EUBIRO-Developers "YouTube Channel"

As a result, a novel structure for the NeuBIRO software was reshaped for Task 8.2.

Figure 8.1 shows the architecture of a complete NeuBIRO data processing as a series of contatenated steps:

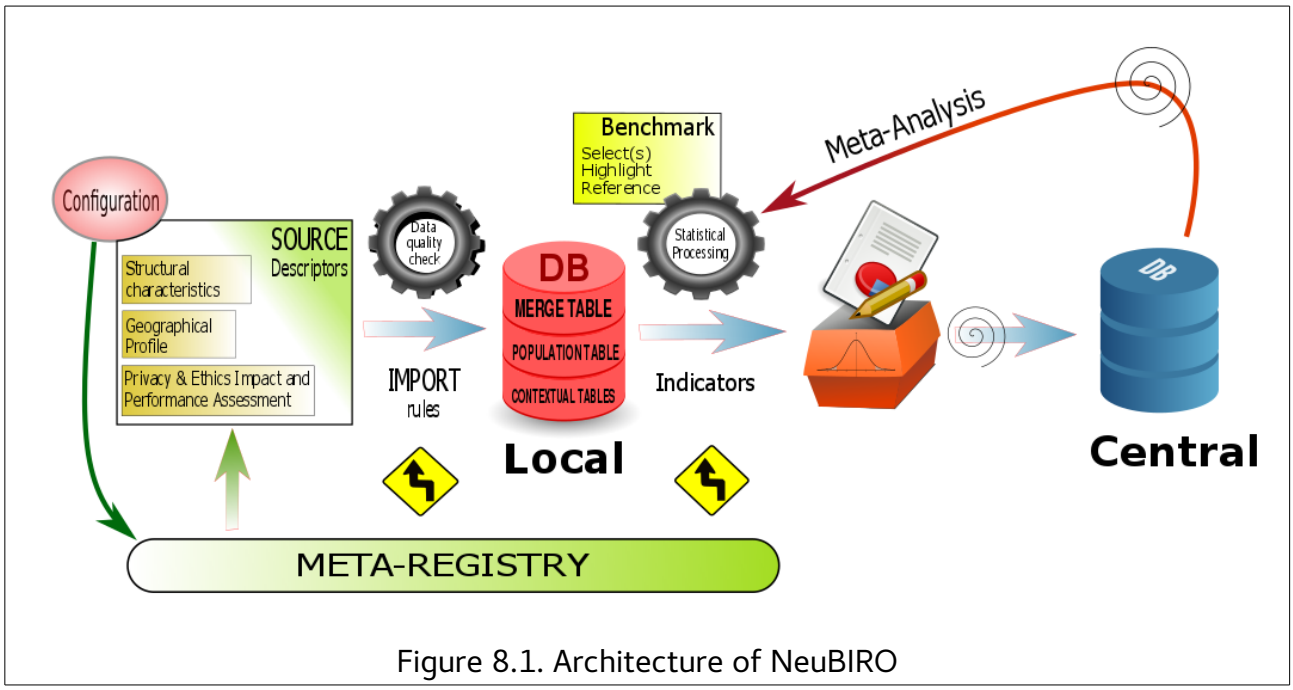

Figure 8.1. Architecture of NeuBIRO

- **Configuration**: each user has unique credentials through which the system recognizes a link to a specific data source profile stored in a specific area of a centralised server

(meta-registry). The registry includes information on the data stored at each site (e.g. type of data available, meta-data of data elements/tables/columns), structural characteristics (e.g. capacity of a setting, workforce involved, etc), geographical characteristics (e.g. location, population data, etc), privacy procedures (privacy and ethics impact and performance assessment questionnaire). In this way, contextual data could be added to the system analysis as supplementary tables, whenever needed. This core component of the system may foster the application of a genuine population-based approach, overcoming some of the problems of bias described above, e.g. attributing a unique catchment area to each data source, using the correct denominators, avoiding duplications of records across data sources, etc. A working example of how the meta-registry can be structured has been developed for the case of diabetes in EUBIROD: http://uat.mydiabetesmyway.scot.nhs.uk/metaregistry. The web interface includes a hierarchical geo-coding structure specifically conceived for BIRO from the Nomenclature of Territorial Units for Statistics (NUTS): http://uat.mydiabetesmyway.scot.nhs.uk/metaregistry/datasource/createcode.php. In this way, a unique catchment area can be assigned to each fundamental entity using the platform at the lowest possible level.

- **Data Import and quality check**: using specifications provided by the meta-registry, the system loads a “csv” (comma-separated values) file in a local database containing all individual records obtained from linked data at each source (eg linked administrative data merged with additional contextual tables eg physician ID, medical group, health district area). This step is only necessary when NeuBIRO is used by a local data custodian at a lowest possible level in the hierarchy of data sources (i.e. where individual records are directly accessible). Data import also allows to pre-process the data and create derived columns that will be added to the database. To simplify the process. the database is recreated at each launch of the import procedure. In this way, any new analysis will start with a "clean" database recreated *ad hoc*. The product of this step is a local database that includes the main linked table (aka “merge table”), plus additional contextual datasets including the population tables for aggregate data from the reference region, the “activity dataset” used as a master index for the actual denominators, etc.
- **Statistical processing**: through specialised software routines, the user can process data for different purposes (data aggregation, statistical analysis, creation of output files, generation of test data to run the program, etc). Statistical routines compute indicators using specifications included in the meta-registry. The "statistical engine" can work in two different modes:
  - **"Local"**: uses individual data stored in the internal database. Pre-processing can be performed on the database to extract subsets of records and/or create aggregate data, so that the analysis would be simplified and be generally faster for statistical routines to operate. The process can finally deliver outputs in different formats (e.g. pdf or dynamic webpages), including the creation of aggregate tables (aka “statistical objects” in the BIRO approach) to be reused at a higher level in the hierarchy of data sources.

- **"Central"**: selecting one or more of the outputs produced in either mode by the modules for statistical analysis of NeuBIRO, it is possible to launch the statistical modules again, producing cumulative results for the entire pool of datasets, in a very flexible *meta-analytical* fashion.

- **Transfer**: the platform includes simple routines to send a package of aggregated data towards a central server, including a set of tables and a descriptor file that would allow the dynamic inclusion into a centralised database that can be continuously used as the resulting common **platform for population-based chronic disease registries**. The platform would become then publicly available for monitoring and benchmarking to a range of stakeholders.

8.3 Implementation of the platform

During the first 18 months of the Bridge Health project, the development of "NeuBIRO" proceeded on a rolling basis, taking into account all suggestions made at the outset by the Study Group.

Software development initially concentrated on the preparation of the environment and the interface, then proceeded in parallel according to the specific background of the programming team. Team work was organized through the revision control software "GIT" (http://git-scm.com/), pointing to a common repository on Gitlab, where all source code was kept accessible to registered users of the Study Group and agreed representatives of Bridge Health coordination (https://gitlab.com/eubirodnetwork/bridge-health).

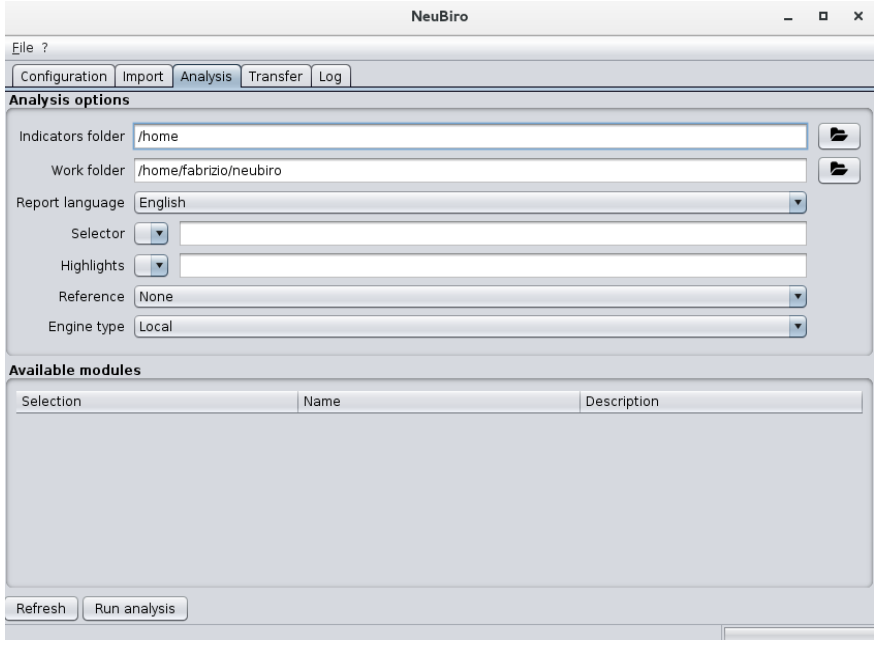

Figure 8.2. NeuBIRO: Interface of the Statistical Engine

The core system was developed using the Java-based Groovy programming language

(Groovy Programming Team 2014). The system provides the user with a simple graphical interface (**Figure 8.2**), specifically designed to be direct and easy to use. The use is guided by a different panel for each function described above, plus a view of execution messages ("log").

The “database engine” of choice adopted for data import and transformation is H2 (http://www.h2database.com/), an agile java-based open source database management system that can also connect to external statistical routines.

The data import specifications were automated using descriptions file whose syntax is represented by an internal DSL based on the Groovy language. Files of this sort (“import.specs”) can be used to pre-process data and apply specific recoding to the original columns (eg. age intervals from age as a continuous measure).

All statistical routines (“statistical engine”) were developed using the "R" system (R Core Team 1013). The scripts are organized in modules that can be loaded by the NeuBIRO, with the option of further pre-processing through the database engine independently from the import routine. The process is regulated by specific scripts written in DSL and Groovy, which can optionally run before any R code in each module.

The following options implemented in the statistical engine are noteworthy:

- adjustment and system factors: targeted parameters may be used to tune the analysis on specific characteristics, classified as follows:
    a) diseases: the report can include specific combinations of target conditions (eg. hypertension, diabetes, etc);
    b) adjustment factors: characteristics (columns e.g. age, sex, etc) that are used as confounders to produce standardised indicators and/or risk adjusted estimates from multivariate models;
    c) system factors: levels used to benchmark results across units and/or sources of variation for multilevel models (e.g. GPs, primary care groups, hospitals, districts, local health authorities, etc.).
- modular composition of the reports: possibility to use only specific modules to tailor reports according to own needs, or add more modules to expand the range of sections available in the report from the same data.
- filtered results (“select”): a report can be produced only on a specific subset of observations selected by applying a specific selection of subjects on the overall dataset (eg. age> 75 years). This option is available either through a guided text entry or via script source for advanced users (R syntax).
- benchmarking against a “reference” subset (internal or external): results included in the report are compared against those produced using the same routines from a reference population (dataset). The reference population can be either "internal" (e.g. subset of the total population selected through a filter) or "external" (user-supplied via external file produced by NeuBIRO using the same choice of modules on a different dataset.
- loop analysis ("batch"): using appropriate commands, the system automatically produces a report for a data subset corresponding to each distinct value in a data column (eg. by

district or local health authority).

- multilanguage support: the system includes general and specific settings enabling the production of reports in different languages (currently implemented only in English and Italian).

Outputs generated by the system have been organized in pdf documents, realised as a compilation of the results from the list of statistical modules selected for the specific analysis. The analysis modules in Docbook format (www.docbook.org/). Docbook is cross-platform, with same functioning on Windows, Linux and Mac operating systems.

Regarding on field tests, so far the new platform has not been used on real datasets from any of the partner sources. Tests have been conducted by the core programming team using a set of statistical routines to produce test indicators on fictitious samples of different size on a mid-end notebook (I7 CPU, 8GB RAM, 700GB HD).

The open source systems adopted have proven to be highly efficient for the purpose, not doing find particular stability problems. Some problems were found for automatic update of some libraries of the R system for which it was necessary to develop special version control routine.

The initial performance shown by the system is quite encouraging in terms of operability of the prototype on very large datasets. The creation of a "virtual" dataset including 2,500,000 records and 45 columns took 20'46" minutes. The dataset was imported by the database engine in a H2 table in 46'03" minutes. A set of indicators were run by the statistical engine in "local mode", resulting into over 30 pages of reports including tables and figures took 29'15" for data transformation, plus 8'45" for data analysis and 54'54" to create outputs for central analysis. Overall, the entire cycle required to create a complex report took approximately 2 hours and 20 minutes on a client machine.

A similar test was made to trial the use of the "central mode" on 4 virtual sites, with an overall dataset of the same dimension. The test demonstrated that the platform can be very fast once the aggregate data are created from local data sources, resulting in the entire cycle to complete in about 15 minutes.

The processing mode inspired by the BIRO approach confirms that a benchmarking tool for Europe can run almost in real time using a centralised database of aggregate data generated by the same tool using standardised definitions.

The initial development resulted in the first release of NeuBIRO made available in October 2016. Several release cycles have been foreseen until the end of the project to implement all suggested changes up to October 2017, when release 0.7 was deployed.

Software complied with the following indications **for chronic-disease reporting**:

- **simplified 'basic reports':** more specific clinical orientation (for local reports) and simplified presentation of national disease indicators for policy making and continuous monitoring. In this way, reports can be compiled at user demand, based on the availability of information from administrative data or clinical registries, for the desired statistical unit.
- **clinical reports made more targeted and dynamic**, using simple association measures

eg relative risks or odds ratios that could better inform clinicians, or by making more specific queries that can be sensible for the clinician

- **further developments:**
  - an import routine that can allow computing all indicators included in BIRO
  - new basic reports for a subset of indicators specified in the manual
  - a scientific paper using data from 2010
  - new data collection from the EUBIROD network
  - test of indicators on costs and socio-economic status

In the final year of implementation, the application has been successfully tested in the field of diabetes, collecting impressions from users and proving that the particular modular format can help using procedures in a nearly autonomous manner.

As the general features have been finally delivered, the application of NeuBIRO is also fit to work on cardiovascular data. Leaders of task 8.1 have been pointed to the particular import specifications that may allow targeting a wide spectrum of domains, from population health to health care, to expand the range of clinical fields covered.

The range of graphical tools has been expanded to cover funnel plots and turnip charts that can show the variability of results across the system in an unbiased manner.

The BIRO approach, until now applied in Europe inly within the community of registers of diabetes, can be further implemented to deal with more complex conditions, including the need for a holistic view of chronic diseases, and the challenge of a shared quality of governance, to support the decentralised implementation of Directive 2011/24 for the supply of cross-border services (European Commission 2011).

In this project, we have also found new ways to comply with the European Directive on the Protection of personal data (European Commission 2016). An innovative solution such as NeuBIRO can provide significant added value to the regulatory functions of the European Union.

Field experience shows that the different levels of the health system shall cooperate not only to provide care, but also to share essential information that can foster policies for the continuous improvement of health services at the organizational, regional, national and European level.

All source code produced with the NeuBIRO is now publicly available free for use, being licensed through the European Public License (EUPL), at the following URL on GitHub:

https://github.com/eubirodnetwork/neubiro

References

- Berwick DM et al (2008). The triple aim: care, health, and cost. Health Aff (Millwood). 27(3):759-69.
- Boyle D et al 2001, Cunningham S, Sullivan F and Morris A, on behalf of the Tayside Regional Diabetes Network, Technology Integration for the provision of population-based equitable patient care: the Tayside Regional Diabetes Network: a brief description, Diabetes Nutrition and Metabolism 2001;14(2), 100-103.
- BIRO Consortium (2009). Best Information through Regional Outcomes: a Shared European Diabetes Information System for Policy and Practice, University of Perugia. Available at: http://www.eubirod.eu/documents/downloads/BIRO_Monograph.pdf.
- Carinci F, Federici MO, Benedetti MM (2006). Diabetes registers and prevention strategies: towards an active use of health information. Diabetes Res Clin Pract. 74 Suppl 2:S215-9.
- Carinci F, Rossi L (2009), The BIRO Statistical Engine. Available at: http://www.biro-project.eu/documents/downloads/D8_1%20Statistical_Engine.pdf.
- Carinci F et al. (2010a). Statistical Materials, University of Perugia. Available at: http://www.eubirod.eu/documents/downloads/D6_1_Statistical_Materials.pdf.
- Carinci F et al (2010b), A Novel International Framework for Privacy-Enhanced Data Processing, Exchange, and Pooled Analysis of Disease Registers: The European BIRO/EUBIROD Projects, Academy Health Research Meeting 2010, Boston, Massachusetts, USA. Available at: http://www.academyhealth.org/files/2010/tuesday/carincif.pdf.
- Carinci F et al (2011). Revision of the European Data Protection Directive: opportunity or threat for public health monitoring? Eur J Public Health. 2011 Dec;21(6):684-5. Available at: http://eurpub.oxfordjournals.org/content/21/6/684.full.pdf+html.
- Carinci F et al. (2015a). Towards actionable international comparisons of health system performance: expert revision of the OECD framework and quality indicators, International Journal for Quality in Health Care, Apr; 27(2):137-46. Available at: http://intqhc.oxfordjournals.org/content/27/2/137.long.
- Carinci F (2015b). Essential levels of health information in Europe: an action plan for a coherent and sustainable infrastructure, Health Policy, 2015 Apr;119(4):530-8.
- Carinci F et al. (2015c). Minutes of the 1st Bridge Health Meeting of the EUBIROD Network, Hotel Archimede, Via Dei Mille 19, Rome, Italy, 23-24th November 2015
- Carinci F et al. (2016). Lower extremity amputation rates in people with diabetes as an indicator of health systems performance. A critical appraisal of the data collection 2000-2011 by the Organization for Economic Cooperation and Development (OECD), Acta Diabetologica, 53(5):825-32. Available at: http://link.springer.com/article/10.1007%2Fs00592-016-0879-4.

- Carstensen B et al (2011). The National Diabetes Register, Scand J Public Health. 39(7 Suppl):58-61. Available at: http://sjp.sagepub.com/content/39/7_suppl/58.long
- Chobot A et al. (2015) Incidence of type 1 diabetes among Polish children ages 0-14 years from 1989-2012. Acta Diabetol. 52(3):483-8.
- Cooper JG et al. (2013). The Norwegian Diabetes Register for Adults – an overview of the first years. Norsk Epidemiologi; 23 (1): 29-34
- Cunningham SG et al (2015a). Defining a European diabetes data dictionary for clinical audit and healthcare delivery: core standards of the EUBIROD project, Methods of Information in Medicine, 2015 Dec 15;55(2).
- Di Iorio CT et al (2009), Privacy impact assessment in the design of transnational public health information systems: the BIRO project, Journal of Medical Ethics, Dec;35(12):753-61, 2009.
- Di Iorio CT, Carinci F (2012). Data Protection and Health Care Information Systems: Where is the Balance?, in "Carlisle G, Whitehouse D, Penny D (Eds), eHealth: Legal, Ethical and Governance Challenges", Springer Verlag.European Commission (2011)
- Di Iorio CT et al (2013a). Health Research and Systems' Governance are at risk: should the right to data protection override health?, J Med Ethics.
- Di Iorio CT et al. (2013b). Cross-border flow of health information: is "privacy by design" enough? Privacy performance assessment in EUBIROD, Eur J Public Health. Apr;23(2):247-53.
- Doggen K et al. (2014) Initiative for Quality Improvement and Epidemiology in Diabetic Foot Clinics (IQED-Foot) Study Group. Implementation of a quality improvement initiative in Belgian diabetic foot clinics: feasibility and initial results. Diabetes Metab Res Rev. 30(5):435-43.
- Dovc K et al. (2014). Improved Metabolic Control in Pediatric Patients with Type 1 Diabetes: A Nationwide Prospective 12-Year Time Trends Analysis. Diabetes Technol Ther. 16(1):33–40.
- European Commission (2009), Article 29 Data Protection Working Party. The Future of Privacy. Joint contribution to the Consultation of the European Commission on the legal framework for the fundamental right to protection of personal data, 02356/09/EN, WP 168- Available at: http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2009/wp168_en.pdf.
- European Commission (2011), Directive 2011/24/EU of the European Parliament and of the Council of 9 March 2011 on the application of patients' rights in cross-border healthcare. Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32011L0024:EN:NOT.
- European Commission (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Available at: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679.

- Jecht M et al. (2015). Charakteristika von Menschen mit einem diabetischen Fußsyndrom. Der Diabetologe; 2:138-43.
- Kern J et al (2012). Public Health ICT Based Surveillance System, Telemedicine and E-Health Services, Policies, and Applications: Advancements and Developments / Rodrigues, Joel J. P. C. ; de la Torre Díez, Isabel ; Sainz de Abajo, Beatriz ; (ur.). Hershey, PA, USA : IGI Global, p.369-396.
- Gini R et al. (2016). Data Extraction and Management in Networks of Observational Health Care Databases for Scientific Research: A Comparison of EU-ADR, OMOP, Mini-Sentinel and MATRICE Strategies. EGEMS (Wash DC). 2016 Feb 8;4(1):1189.
- Groovy Programming Team (2014), Groovy. A dynamic language for Java platform. Available at: http://groovy.codehaus.org/.
- Häkkinen U et al. (2013) Health care performance comparison using a disease-based approach: the EuroHOPE project. Health Policy. 112(1-2):100-9.
- McAlpine R (2009), communication to the 1st EUBIROD annual meeting, Kuwait City, Kuwait, 3rd May 2009. Available at: http://www.eubirod.eu/academy/first_course/lectures/Tayside_mcalpine.html.
- Morris AD et al 1997, The diabetes audit and research in Tayside Scotland (DARTS) study: electronic record linkage to create a diabetes register. BMJ; 315:524–528, available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2127363/pdf/9329309.pdf.
- PARENT Consortium (2014), PARENT Joint Action Guidelines. Available at: http://patientregistries.eu/documents/10184/14613/Methodological+guidelines+and+recommendations+for+efficient+and+rational+governance+of+patient+registries/5bf3ea46-26ca-4d91-b459-47dce18a8846.
- Pildava S et al (2014). The mortality of patients with diabetes mellitus in Latvia 2000–2012, Lithuanian University of Health Sciences journal Medicine.
- R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org/.
- Raymond, ES (1999). The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary. O'Reilly Media.
- United Nations (2011), Political Declaration of the UN High-Level Meeting on the Prevention and Control of Non-communicable Diseases (NCDs). Available at: :http://www.who.int/entity/nmh/events/un_ncd_summit2011/political_declaration_en.pdf?ua=1.
- WHO (2015), Integrated Surveillance of noncommunicable diseases. Available at: http://www.euro.who.int/en/health-topics/noncommunicable-diseases/integrated-surveillance-of-noncommunicable-diseases-incd
- Van der Wilk EA, Achterberg PW, Verschuuren M, Harbers MM (2012). EU reflection Process on Chronic Disease. Results of the Stakeholder Consultation, RIVM, Bilthoven, The Netherlands. Available at:

http://ec.europa.eu/health/major_chronic_diseases/docs/reflection_process_final_report_stakeholder_consultation_2012_en.pdf